Graduate Texts in Mathematics 176

Springer
New York
Berlin
Heidelberg
Barcelona
Budapest
Hong Kong
London
Milan
Paris
Santa Clara
Singapore
Tokyo

Graduate Texts in Mathematics

1 TAKEUTI/ZARING. Introduction to Axiomatic Set Theory. 2nd ed.
2 OXTOBY. Measure and Category. 2nd ed.
3 SCHAEFER. Topological Vector Spaces.
4 HILTON/STAMMBACH. A Course in Homological Algebra. 2nd ed.
5 MAC LANE. Categories for the Working Mathematician.
6 HUGHES/PIPER. Projective Planes.
7 SERRE. A Course in Arithmetic.
8 TAKEUTI/ZARING. Axiomatic Set Theory.
9 HUMPHREYS. Introduction to Lie Algebras and Representation Theory.
10 COHEN. A Course in Simple Homotopy Theory.
11 CONWAY. Functions of One Complex Variable I. 2nd ed.
12 BEALS. Advanced Mathematical Analysis.
13 ANDERSON/FULLER. Rings and Categories of Modules. 2nd ed.
14 GOLUBITSKY/GUILLEMIN. Stable Mappings and Their Singularities.
15 BERBERIAN. Lectures in Functional Analysis and Operator Theory.
16 WINTER. The Structure of Fields.
17 ROSENBLATT. Random Processes. 2nd ed.
18 HALMOS. Measure Theory.
19 HALMOS. A Hilbert Space Problem Book. 2nd ed.
20 HUSEMOLLER. Fibre Bundles. 3rd ed.
21 HUMPHREYS. Linear Algebraic Groups.
22 BARNES/MACK. An Algebraic Introduction to Mathematical Logic.
23 GREUB. Linear Algebra. 4th ed.
24 HOLMES. Geometric Functional Analysis and Its Applications.
25 HEWITT/STROMBERG. Real and Abstract Analysis.
26 MANES. Algebraic Theories.
27 KELLEY. General Topology.
28 ZARISKI/SAMUEL. Commutative Algebra. Vol.I.
29 ZARISKI/SAMUEL. Commutative Algebra. Vol.II.
30 JACOBSON. Lectures in Abstract Algebra I. Basic Concepts.
31 JACOBSON. Lectures in Abstract Algebra II. Linear Algebra.
32 JACOBSON. Lectures in Abstract Algebra III. Theory of Fields and Galois Theory.
33 HIRSCH. Differential Topology.
34 SPITZER. Principles of Random Walk. 2nd ed.
35 WERMER. Banach Algebras and Several Complex Variables. 2nd ed.
36 KELLEY/NAMIOKA et al. Linear Topological Spaces.
37 MONK. Mathematical Logic.
38 GRAUERT/FRITZSCHE. Several Complex Variables.
39 ARVESON. An Invitation to C^*-Algebras.
40 KEMENY/SNELL/KNAPP. Denumerable Markov Chains. 2nd ed.
41 APOSTOL. Modular Functions and Dirichlet Series in Number Theory. 2nd ed.
42 SERRE. Linear Representations of Finite Groups.
43 GILLMAN/JERISON. Rings of Continuous Functions.
44 KENDIG. Elementary Algebraic Geometry.
45 LOÈVE. Probability Theory I. 4th ed.
46 LOÈVE. Probability Theory II. 4th ed.
47 MOISE. Geometric Topology in Dimensions 2 and 3.
48 SACHS/WU. General Relativity for Mathematicians.
49 GRUENBERG/WEIR. Linear Geometry. 2nd ed.
50 EDWARDS. Fermat's Last Theorem.
51 KLINGENBERG. A Course in Differential Geometry.
52 HARTSHORNE. Algebraic Geometry.
53 MANIN. A Course in Mathematical Logic.
54 GRAVER/WATKINS. Combinatorics with Emphasis on the Theory of Graphs.
55 BROWN/PEARCY. Introduction to Operator Theory I: Elements of Functional Analysis.
56 MASSEY. Algebraic Topology: An Introduction.
57 CROWELL/FOX. Introduction to Knot Theory.
58 KOBLITZ. p-adic Numbers, p-adic Analysis, and Zeta-Functions. 2nd ed.
59 LANG. Cyclotomic Fields.
60 ARNOLD. Mathematical Methods in Classical Mechanics. 2nd ed.

continued after index

John M. Lee

Riemannian Manifolds

An Introduction to Curvature

With 88 Illustrations

Springer

John M. Lee
Department of Mathematics
University of Washington
Seattle, WA 98195-4350
USA

Mathematics Subject Classification (1991): 53-01, 53C20

Library of Congress Cataloging-in-Publication Data
Lee, John M., 1950–
Reimannian manifolds : an introduction to curvature / John M. Lee.
p. cm. — (Graduate texts in mathematics ; 176)
Includes index.

1. Reimannian manifolds. I. Title. II. Series.
QA649.L397 1997
516.3′73—dc21 97-14537

Printed on acid-free paper.

Production managed by Lesley Poliner; manufacturing supervised by Joe Quatela.
Photocomposed pages prepared from the author's TEX files.

9 8 7 6 5 4 3 2 1

ISBN 978-0-387-98322-6 ISBN 978-0-387-22726-9 (eBook)
DOI 10.1007/978-0-387-22726-9

To my family:
Pm, Nathan, and Jeremy Weizenbaum

Preface

This book is designed as a textbook for a one-quarter or one-semester graduate course on Riemannian geometry, for students who are familiar with topological and differentiable manifolds. It focuses on developing an intimate acquaintance with the geometric meaning of curvature. In so doing, it introduces and demonstrates the uses of all the main technical tools needed for a careful study of Riemannian manifolds.

I have selected a set of topics that can reasonably be covered in ten to fifteen weeks, instead of making any attempt to provide an encyclopedic treatment of the subject. The book begins with a careful treatment of the machinery of metrics, connections, and geodesics, without which one cannot claim to be doing Riemannian geometry. It then introduces the Riemann curvature tensor, and quickly moves on to submanifold theory in order to give the curvature tensor a concrete quantitative interpretation. From then on, all efforts are bent toward proving the four most fundamental theorems relating curvature and topology: the Gauss–Bonnet theorem (expressing the total curvature of a surface in terms of its topological type), the Cartan–Hadamard theorem (restricting the topology of manifolds of nonpositive curvature), Bonnet's theorem (giving analogous restrictions on manifolds of strictly positive curvature), and a special case of the Cartan–Ambrose–Hicks theorem (characterizing manifolds of constant curvature).

Many other results and techniques might reasonably claim a place in an introductory Riemannian geometry course, but could not be included due to time constraints. In particular, I do not treat the Rauch comparison theorem, the Morse index theorem, Toponogov's theorem, or their important applications such as the sphere theorem, except to mention some of them

in passing; and I do not touch on the Laplace–Beltrami operator or Hodge theory, or indeed any of the multitude of deep and exciting applications of partial differential equations to Riemannian geometry. These important topics are for other, more advanced courses.

The libraries already contain a wealth of superb reference books on Riemannian geometry, which the interested reader can consult for a deeper treatment of the topics introduced here, or can use to explore the more esoteric aspects of the subject. Some of my favorites are the elegant introduction to comparison theory by Jeff Cheeger and David Ebin [CE75] (which has sadly been out of print for a number of years); Manfredo do Carmo's much more leisurely treatment of the same material and more [dC92]; Barrett O'Neill's beautifully integrated introduction to pseudo-Riemannian and Riemannian geometry [O'N83]; Isaac Chavel's masterful recent introductory text [Cha93], which starts with the foundations of the subject and quickly takes the reader deep into research territory; Michael Spivak's classic tome [Spi79], which can be used as a textbook if plenty of time is available, or can provide enjoyable bedtime reading; and, of course, the "Encyclopaedia Britannica" of differential geometry books, *Foundations of Differential Geometry* by Kobayashi and Nomizu [KN63]. At the other end of the spectrum, Frank Morgan's delightful little book [Mor93] touches on most of the important ideas in an intuitive and informal way with lots of pictures—I enthusiastically recommend it as a prelude to this book.

It is not my purpose to replace any of these. Instead, it is my hope that this book will fill a niche in the literature by presenting a selective introduction to the main ideas of the subject in an easily accessible way. The selection is small enough to fit into a single course, but broad enough, I hope, to provide any novice with a firm foundation from which to pursue research or develop applications in Riemannian geometry and other fields that use its tools.

This book is written under the assumption that the student already knows the fundamentals of the theory of topological and differential manifolds, as treated, for example, in [Mas67, chapters 1–5] and [Boo86, chapters 1–6]. In particular, the student should be conversant with the fundamental group, covering spaces, the classification of compact surfaces, topological and smooth manifolds, immersions and submersions, vector fields and flows, Lie brackets and Lie derivatives, the Frobenius theorem, tensors, differential forms, Stokes's theorem, and elementary properties of Lie groups. On the other hand, I do not assume any previous acquaintance with Riemannian metrics, or even with the classical theory of curves and surfaces in $\mathbf{R}^3$. (In this subject, anything proved before 1950 can be considered "classical.") Although at one time it might have been reasonable to expect most mathematics students to have studied surface theory as undergraduates, few current North American undergraduate math majors see any differen-

tial geometry. Thus the fundamentals of the geometry of surfaces, including a proof of the Gauss–Bonnet theorem, are worked out from scratch here.

The book begins with a nonrigorous overview of the subject in Chapter 1, designed to introduce some of the intuitions underlying the notion of curvature and to link them with elementary geometric ideas the student has seen before. This is followed in Chapter 2 by a brief review of some background material on tensors, manifolds, and vector bundles, included because these are the basic tools used throughout the book and because often they are not covered in quite enough detail in elementary courses on manifolds. Chapter 3 begins the course proper, with definitions of Riemannian metrics and some of their attendant flora and fauna. The end of the chapter describes the constant curvature "model spaces" of Riemannian geometry, with a great deal of detailed computation. These models form a sort of *leitmotif* throughout the text, and serve as illustrations and testbeds for the abstract theory as it is developed. Other important classes of examples are developed in the problems at the ends of the chapters, particularly invariant metrics on Lie groups and Riemannian submersions.

Chapter 4 introduces connections. In order to isolate the important properties of connections that are independent of the metric, as well as to lay the groundwork for their further study in such arenas as the Chern–Weil theory of characteristic classes and the Donaldson and Seiberg–Witten theories of gauge fields, connections are defined first on arbitrary vector bundles. This has the further advantage of making it easy to define the induced connections on tensor bundles. Chapter 5 investigates connections in the context of Riemannian manifolds, developing the Riemannian connection, its geodesics, the exponential map, and normal coordinates. Chapter 6 continues the study of geodesics, focusing on their distance-minimizing properties. First, some elementary ideas from the calculus of variations are introduced to prove that every distance-minimizing curve is a geodesic. Then the Gauss lemma is used to prove the (partial) converse—that every geodesic is locally minimizing. Because the Gauss lemma also gives an easy proof that minimizing curves are geodesics, the calculus-of-variations methods are not strictly necessary at this point; they are included to facilitate their use later in comparison theorems.

Chapter 7 unveils the first fully general definition of curvature. The curvature tensor is motivated initially by the question of whether all Riemannian metrics are locally equivalent, and by the failure of parallel translation to be path-independent as an obstruction to local equivalence. This leads naturally to a qualitative interpretation of curvature as the obstruction to flatness (local equivalence to Euclidean space). Chapter 8 departs somewhat from the traditional order of presentation, by investigating submanifold theory immediately after introducing the curvature tensor, so as to define sectional curvatures and give the curvature a more quantitative geometric interpretation.

The last three chapters are devoted to the most important elementary global theorems relating geometry to topology. Chapter 9 gives a simple moving-frames proof of the Gauss–Bonnet theorem, complete with a careful treatment of Hopf's rotation angle theorem (the *Umlaufsatz*). Chapter 10 is largely of a technical nature, covering Jacobi fields, conjugate points, the second variation formula, and the index form for later use in comparison theorems. Finally in Chapter 11 comes the *dénouement*—proofs of some of the "big" global theorems illustrating the ways in which curvature and topology affect each other: the Cartan–Hadamard theorem, Bonnet's theorem (and its generalization, Myers's theorem), and Cartan's characterization of manifolds of constant curvature.

The book contains many questions for the reader, which deserve special mention. They fall into two categories: "exercises," which are integrated into the text, and "problems," grouped at the end of each chapter. Both are essential to a full understanding of the material, but they are of somewhat different character and serve different purposes.

The exercises include some background material that the student should have seen already in an earlier course, some proofs that fill in the gaps from the text, some simple but illuminating examples, and some intermediate results that are used in the text or the problems. They are, in general, elementary, but they are *not optional*—indeed, they are integral to the continuity of the text. They are chosen and timed so as to give the reader opportunities to pause and think over the material that has just been introduced, to practice working with the definitions, and to develop skills that are used later in the book. I recommend strongly that students stop and do each exercise as it occurs in the text before going any further.

The problems that conclude the chapters are generally more difficult than the exercises, some of them considerably so, and should be considered a central part of the book by any student who is serious about learning the subject. They not only introduce new material not covered in the body of the text, but they also provide the student with indispensable practice in using the techniques explained in the text, both for doing computations and for proving theorems. If more than a semester is available, the instructor might want to present some of these problems in class.

Acknowledgments: I owe an unpayable debt to the authors of the many Riemannian geometry books I have used and cherished over the years, especially the ones mentioned above—I have done little more than rearrange their ideas into a form that seems handy for teaching. Beyond that, I would like to thank my Ph.D. advisor, Richard Melrose, who many years ago introduced me to differential geometry in his eccentric but thoroughly enlightening way; Judith Arms, who, as a fellow teacher of Riemannian geometry at the University of Washington, helped brainstorm about the "ideal contents" of this course; all my graduate students at the University

of Washington who have suffered with amazing grace through the flawed early drafts of this book, especially Jed Mihalisin, who gave the manuscript a meticulous reading from a user's viewpoint and came up with numerous valuable suggestions; and Ina Lindemann of Springer-Verlag, who encouraged me to turn my lecture notes into a book and gave me free rein in deciding on its shape and contents. And of course my wife, Pm Weizenbaum, who contributed professional editing help as well as the loving support and encouragement I need to keep at this day after day.

Contents

1
What Is Curvature?

If you've just completed an introductory course on differential geometry, you might be wondering where the geometry went. In most people's experience, geometry is concerned with properties such as distances, lengths, angles, areas, volumes, and curvature. These concepts, however, are barely mentioned in typical beginning graduate courses in differential geometry; instead, such courses are concerned with smooth structures, flows, tensors, and differential forms.

The purpose of this book is to introduce the theory of *Riemannian manifolds:* these are smooth manifolds equipped with Riemannian metrics (smoothly varying choices of inner products on tangent spaces), which allow one to measure geometric quantities such as distances and angles. This is the branch of modern differential geometry in which "geometric" ideas, in the familiar sense of the word, come to the fore. It is the direct descendant of Euclid's plane and solid geometry, by way of Gauss's theory of curved surfaces in space, and it is a dynamic subject of contemporary research.

The central unifying theme in current Riemannian geometry research is the notion of curvature and its relation to topology. This book is designed to help you develop both the tools and the intuition you will need for an in-depth exploration of curvature in the Riemannian setting. Unfortunately, as you will soon discover, an adequate development of curvature in an arbitrary number of dimensions requires a great deal of technical machinery, making it easy to lose sight of the underlying geometric content. To put the subject in perspective, therefore, let's begin by asking some very basic questions: What is curvature? What are the important theorems about it?

In this chapter, we explore these and related questions in an informal way, without proofs. In the next chapter, we review some basic material about tensors, manifolds, and vector bundles that is used throughout the book. The "official" treatment of the subject begins in Chapter 3.

The Euclidean Plane

To get a sense of the kinds of questions Riemannian geometers address and where these questions came from, let's look back at the very roots of our subject. The treatment of geometry as a mathematical subject began with Euclidean plane geometry, which you studied in school. Its elements are points, lines, distances, angles, and areas. Here are a couple of typical theorems:

Theorem 1.1. (SSS) *Two Euclidean triangles are congruent if and only if the lengths of their corresponding sides are equal.*

Theorem 1.2. (Angle-Sum Theorem) *The sum of the interior angles of a Euclidean triangle is π.*

As trivial as they seem, these two theorems serve to illustrate two major types of results that permeate the study of geometry; in this book, we call them "classification theorems" and "local-global theorems."

The SSS (Side-Side-Side) theorem is a *classification theorem.* Such a theorem tells us that to determine whether two mathematical objects are equivalent (under some appropriate equivalence relation), we need only compare a small (or at least finite!) number of computable invariants. In this case the equivalence relation is congruence—equivalence under the group of rigid motions of the plane—and the invariants are the three side lengths.

The angle-sum theorem is of a different sort. It relates a local geometric property (angle measure) to a global property (that of being a three-sided polygon or triangle). Most of the theorems we study in this book are of this type, which, for lack of a better name, we call *local-global theorems.*

After proving the basic facts about points and lines and the figures constructed directly from them, one can go on to study other figures derived from the basic elements, such as circles. Two typical results about circles are given below; the first is a classification theorem, while the second is a local-global theorem. (It may not be obvious at this point why we consider the second to be a local-global theorem, but it will become clearer soon.)

Theorem 1.3. (Circle Classification Theorem) *Two circles in the Euclidean plane are congruent if and only if they have the same radius.*

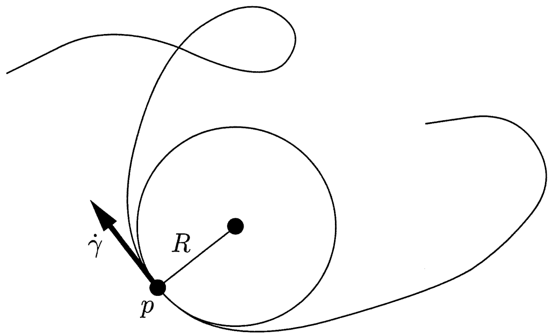


FIGURE 1.1. Osculating circle.

Theorem 1.4. (Circumference Theorem) *The circumference of a Euclidean circle of radius R is $2\pi R$.*

If you want to continue your study of plane geometry beyond figures constructed from lines and circles, sooner or later you will have to come to terms with other curves in the plane. An arbitrary curve cannot be completely described by one or two numbers such as length or radius; instead, the basic invariant is curvature, which is defined using calculus and is a function of position on the curve.

Formally, the *curvature* of a plane curve γ is defined to be $\kappa(t) := |\ddot{\gamma}(t)|$, the length of the acceleration vector, when γ is given a unit speed parametrization. (Here and throughout this book, we think of curves as parametrized by a real variable t, with a dot representing a derivative with respect to t.) Geometrically, the curvature has the following interpretation. Given a point $p = \gamma(t)$, there are many circles tangent to γ at p—namely, those circles that have a parametric representation whose velocity vector at p is the same as that of γ, or, equivalently, all the circles whose centers lie on the line orthogonal to $\dot{\gamma}$ at p. Among these parametrized circles, there is exactly one whose acceleration vector at p is the same as that of γ; it is called the *osculating circle* (Figure 1.1). (If the acceleration of γ is zero, replace the osculating circle by a straight line, thought of as a "circle with infinite radius.") The curvature is then $\kappa(t) = 1/R$, where R is the radius of the osculating circle. The larger the curvature, the greater the acceleration and the smaller the osculating circle, and therefore the faster the curve is turning. A circle of radius R obviously has constant curvature $\kappa \equiv 1/R$, while a straight line has curvature zero.

It is often convenient for some purposes to extend the definition of the curvature, allowing it to take on both positive and negative values. This is done by choosing a unit normal vector field N along the curve, and assigning the curvature a positive sign if the curve is turning toward the

chosen normal or a negative sign if it is turning away from it. The resulting function κ_N along the curve is then called the *signed curvature.*

Here are two typical theorems about plane curves:

Theorem 1.5. (Plane Curve Classification Theorem) *Suppose γ and $\tilde{\gamma}\colon [a,b] \to \mathbf{R}^2$ are smooth, unit speed plane curves with unit normal vector fields N and $\widetilde{N}$, and $\kappa_N(t)$, $\kappa_{\widetilde{N}}(t)$ represent the signed curvatures at $\gamma(t)$ and $\tilde{\gamma}(t)$, respectively. Then γ and $\tilde{\gamma}$ are congruent (by a direction-preserving congruence) if and only if $\kappa_N(t) = \kappa_{\widetilde{N}}(t)$ for all $t \in [a,b]$.*

Theorem 1.6. (Total Curvature Theorem) *If $\gamma\colon [a,b] \to \mathbf{R}^2$ is a unit speed simple closed curve such that $\dot{\gamma}(a) = \dot{\gamma}(b)$, and N is the inward-pointing normal, then*

$$\int_a^b \kappa_N(t)\, dt = 2\pi.$$

The first of these is a classification theorem, as its name suggests. The second is a local-global theorem, since it relates the local property of curvature to the global (topological) property of being a simple closed curve. The second will be derived as a consequence of a more general result in Chapter 9; the proof of the first is left to Problem 9-6.

It is interesting to note that when we specialize to circles, these theorems reduce to the two theorems about circles above: Theorem 1.5 says that two circles are congruent if and only if they have the same curvature, while Theorem 1.6 says that if a circle has curvature κ and circumference C, then $\kappa C = 2\pi$. It is easy to see that these two results are equivalent to Theorems 1.3 and 1.4. This is why it makes sense to consider the circumference theorem as a local-global theorem.

Surfaces in Space

The next step in generalizing Euclidean geometry is to start working in three dimensions. After investigating the basic elements of "solid geometry"—points, lines, planes, distances, angles, areas, volumes—and the objects derived from them, such as polyhedra and spheres, one is led to study more general curved surfaces in space (2-dimensional embedded submanifolds of $\mathbf{R}^3$, in the language of differential geometry). The basic invariant in this setting is again curvature, but it's a bit more complicated than for plane curves, because a surface can curve differently in different directions.

The curvature of a surface in space is described by two numbers at each point, called the principal curvatures. We define them formally in Chapter 8, but here's an informal recipe for computing them. Suppose S is a surface in $\mathbf{R}^3$, p is a point in S, and N is a unit normal vector to S at p.

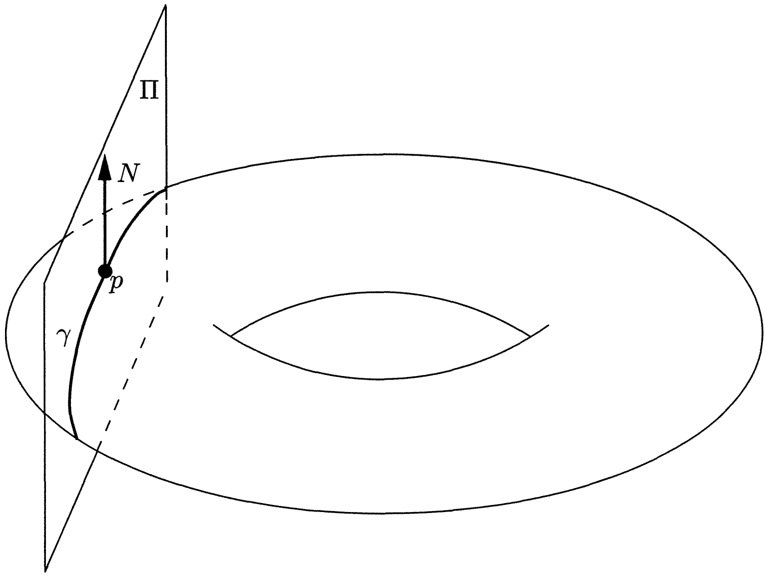


FIGURE 1.2. Computing principal curvatures.

1. Choose a plane Π through p that contains N. The intersection of Π with S is then a plane curve $\gamma \subset \Pi$ passing through p (Figure 1.2).

2. Compute the signed curvature κ_N of γ at p with respect to the chosen unit normal N.

3. Repeat this for *all* normal planes Π. The *principal curvatures* of S at p, denoted κ_1 and κ_2, are defined to be the minimum and maximum signed curvatures so obtained.

Although the principal curvatures give us a lot of information about the geometry of S, they do not directly address a question that turns out to be of paramount importance in Riemannian geometry: Which properties of a surface are intrinsic? Roughly speaking, intrinsic properties are those that could in principle be measured or determined by a 2-dimensional being living entirely within the surface. More precisely, a property of surfaces in $\mathbf{R}^3$ is called *intrinsic* if it is preserved by *isometries* (maps from one surface to another that preserve lengths of curves).

To see that the principal curvatures are not intrinsic, consider the following two embedded surfaces S_1 and S_2 in $\mathbf{R}^3$ (Figures 1.3 and 1.4). S_1 is the portion of the xy-plane where $0 < y < \pi$, and S_2 is the half-cylinder $\{(x,y,z) : y^2 + z^2 = 1, z > 0\}$. If we follow the recipe above for computing principal curvatures (using, say, the downward-pointing unit normal), we find that, since all planes intersect S_1 in straight lines, the principal cur-

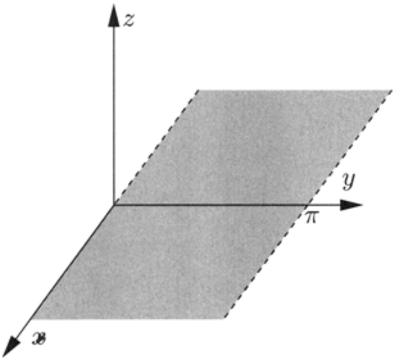


FIGURE 1.3. S_1.

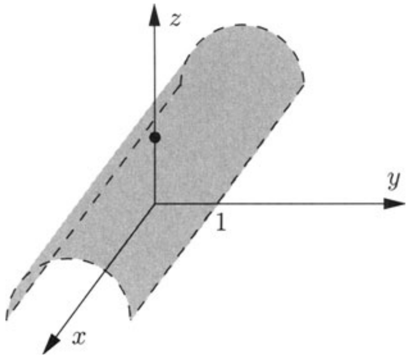


FIGURE 1.4. S_2.

vatures of S_1 are $\kappa_1 = \kappa_2 = 0$. On the other hand, it is not hard to see that the principal curvatures of S_2 are $\kappa_1 = 0$ and $\kappa_2 = 1$. However, the map taking $(x, y, 0)$ to $(x, \cos y, \sin y)$ is a diffeomorphism between S_1 and S_2 that preserves lengths of curves, and is thus an isometry.

Even though the principal curvatures are not intrinsic, Gauss made the surprising discovery in 1827 [Gau65] (see also [Spi79, volume 2] for an excellent annotated version of Gauss's paper) that a particular combination of them is intrinsic. He found a proof that the product $K = \kappa_1\kappa_2$, now called the *Gaussian curvature,* is intrinsic. He thought this result was so amazing that he named it *Theorema Egregium*, which in colloquial American English can be translated roughly as "Totally Awesome Theorem." We prove it in Chapter 8.

To get a feeling for what Gaussian curvature tells us about surfaces, let's look at a few examples. Simplest of all is the plane, which, as we have seen, has both principal curvatures equal to zero and therefore has constant Gaussian curvature equal to zero. The half-cylinder described above also has $K = \kappa_1\kappa_2 = 0 \cdot 1 = 0$. Another simple example is a sphere of radius R. Any normal plane intersects the sphere in great circles, which have radius R and therefore curvature $\pm 1/R$ (with the sign depending on whether we choose the outward-pointing or inward-pointing normal). Thus the principal curvatures are both equal to $\pm 1/R$, and the Gaussian curvature is $\kappa_1\kappa_2 = 1/R^2$. Note that while the signs of the principal curvatures depend on the choice of unit normal, the Gaussian curvature does not: it is always positive on the sphere.

Similarly, any surface that is "bowl-shaped" or "dome-shaped" has positive Gaussian curvature (Figure 1.5), because the two principal curvatures always have the same sign, regardless of which normal is chosen. On the other hand, the Gaussian curvature of any surface that is "saddle-shaped"

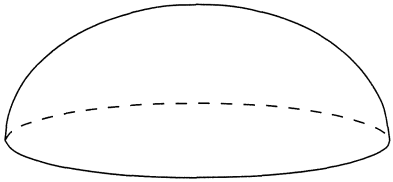


FIGURE 1.5. $K > 0$.

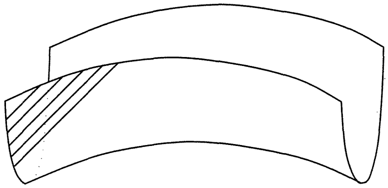


FIGURE 1.6. $K < 0$.

is negative (Figure 1.6), because the principal curvatures are of opposite signs.

The model spaces of surface theory are the surfaces with constant Gaussian curvature. We have already seen two of them: the Euclidean plane $\mathbf{R}^2$ ($K = 0$), and the sphere of radius R ($K = 1/R^2$). The third model is a surface of constant negative curvature, which is not so easy to visualize because it cannot be realized globally as an embedded surface in $\mathbf{R}^3$. Nonetheless, for completeness, let's just mention that the upper half-plane $\{(x, y) : y > 0\}$ with the Riemannian metric $g = R^2 y^{-2}(dx^2 + dy^2)$ has constant negative Gaussian curvature $K = -1/R^2$. In the special case $R = 1$ (so $K = -1$), this is called the *hyperbolic plane.*

Surface theory is a highly developed branch of geometry. Of all its results, two—a classification theorem and a local-global theorem—are universally acknowledged as the most important.

Theorem 1.7. (Uniformization Theorem) *Every connected 2-manifold is diffeomorphic to a quotient of one of the three constant curvature model surfaces listed above by a discrete group of isometries acting freely and properly discontinuously. Therefore, every connected 2-manifold has a complete Riemannian metric with constant Gaussian curvature.*

Theorem 1.8. (Gauss–Bonnet Theorem) *Let S be an oriented compact 2-manifold with a Riemannian metric. Then*

$$\int_S K\, dA = 2\pi\chi(S),$$

where $\chi(S)$ is the Euler characteristic of S (which is equal to 2 if S is the sphere, 0 if it is the torus, and $2 - 2g$ if it is an orientable surface of genus g).

The uniformization theorem is a classification theorem, because it replaces the problem of classifying surfaces with that of classifying discrete groups of isometries of the models. The latter problem is not easy by any means, but it sheds a great deal of new light on the topology of surfaces nonetheless. Although stated here as a geometric-topological result, the uniformization theorem is usually stated somewhat differently and proved

using complex analysis; we do not give a proof here. If you are familiar with complex analysis and the complex version of the uniformization theorem, it will be an enlightening exercise after you have finished this book to prove that the complex version of the theorem is equivalent to the one stated here.

The Gauss–Bonnet theorem, on the other hand, is purely a theorem of differential geometry, arguably the most fundamental and important one of all. We go through a detailed proof in Chapter 9.

Taken together, these theorems place strong restrictions on the types of metrics that can occur on a given surface. For example, one consequence of the Gauss–Bonnet theorem is that the only compact, connected, orientable surface that admits a metric of strictly positive Gaussian curvature is the sphere. On the other hand, if a compact, connected, orientable surface has nonpositive Gaussian curvature, the Gauss–Bonnet theorem forces its genus to be at least 1, and then the uniformization theorem tells us that its universal covering space is topologically equivalent to the plane.

Curvature in Higher Dimensions

We end our survey of the basic ideas of geometry by mentioning briefly how curvature appears in higher dimensions. Suppose M is an n-dimensional manifold equipped with a Riemannian metric g. As with surfaces, the basic geometric invariant is curvature, but curvature becomes a much more complicated quantity in higher dimensions because a manifold may curve in so many directions.

The first problem we must contend with is that, in general, Riemannian manifolds are not presented to us as embedded submanifolds of Euclidean space. Therefore, we must abandon the idea of cutting out curves by intersecting our manifold with planes, as we did when defining the principal curvatures of a surface in $\mathbf{R}^3$. Instead, we need a more intrinsic way of sweeping out submanifolds. Fortunately, *geodesics*—curves that are the shortest paths between nearby points—are ready-made tools for this and many other purposes in Riemannian geometry. Examples are straight lines in Euclidean space and great circles on a sphere.

The most fundamental fact about geodesics, which we prove in Chapter 4, is that given any point $p \in M$ and any vector V tangent to M at p, there is a unique geodesic starting at p with initial tangent vector V.

Here is a brief recipe for computing some curvatures at a point $p \in M$:

1. Pick a 2-dimensional subspace Π of the tangent space to M at p.

2. Look at all the geodesics through p whose initial tangent vectors lie in the selected plane Π. It turns out that near p these sweep out a certain 2-dimensional submanifold S_Π of M, which inherits a Riemannian metric from M.

3. Compute the Gaussian curvature of S_Π at p, which the *Theorema Egregium* tells us can be computed from its Riemannian metric. This gives a number, denoted $K(\Pi)$, called the *sectional curvature* of M at p associated with the plane Π.

Thus the "curvature" of M at p has to be interpreted as a map

$$K\colon \{\text{2-planes in } T_pM\} \to \mathbf{R}.$$

Again we have three constant (sectional) curvature model spaces: $\mathbf{R}^n$ with its Euclidean metric (for which $K \equiv 0$); the n-sphere $\mathbf{S}^n_R$ of radius R, with the Riemannian metric inherited from $\mathbf{R}^{n+1}$ ($K \equiv 1/R^2$); and hyperbolic space $\mathbf{H}^n_R$ of radius R, which is the upper half-space $\{x \in \mathbf{R}^n : x^n > 0\}$ with the metric $h_R := R^2(x^n)^{-2}\sum(dx^i)^2$ ($K \equiv -1/R^2$). Unfortunately, however, there is as yet no satisfactory uniformization theorem for Riemannian manifolds in higher dimensions. In particular, it is definitely not true that every manifold possesses a metric of constant sectional curvature. In fact, the constant curvature metrics can all be described rather explicitly by the following classification theorem.

Theorem 1.9. (Classification of Constant Curvature Metrics) *A complete, connected Riemannian manifold M with constant sectional curvature is isometric to $\widetilde{M}/\Gamma$, where $\widetilde{M}$ is one of the constant curvature model spaces $\mathbf{R}^n$, $\mathbf{S}^n_R$, or $\mathbf{H}^n_R$, and Γ is a discrete group of isometries of $\widetilde{M}$, isomorphic to $\pi_1(M)$, and acting freely and properly discontinuously on $\widetilde{M}$.*

On the other hand, there are a number of powerful local-global theorems, which can be thought of as generalizations of the Gauss–Bonnet theorem in various directions. They are consequences of the fact that positive curvature makes geodesics converge, while negative curvature forces them to spread out. Here are two of the most important such theorems:

Theorem 1.10. (Cartan–Hadamard) *Suppose M is a complete, connected Riemannian n-manifold with all sectional curvatures less than or equal to zero. Then the universal covering space of M is diffeomorphic to $\mathbf{R}^n$.*

Theorem 1.11. (Bonnet) *Suppose M is a complete, connected Riemannian manifold with all sectional curvatures bounded below by a positive constant. Then M is compact and has a finite fundamental group.*

Looking back at the remarks concluding the section on surfaces above, you can see that these last three theorems generalize some of the consequences of the uniformization and Gauss–Bonnet theorems, although not their full strength. It is the primary goal of this book to prove Theorems

1.9, 1.10, and 1.11; it is a primary goal of current research in Riemannian geometry to improve upon them and further generalize the results of surface theory to higher dimensions.

2

Review of Tensors, Manifolds, and Vector Bundles

Most of the technical machinery of Riemannian geometry is built up using tensors; indeed, Riemannian metrics themselves are tensors. Thus we begin by reviewing the basic definitions and properties of tensors on a finite-dimensional vector space. When we put together spaces of tensors on a manifold, we obtain a particularly useful type of geometric structure called a "vector bundle," which plays an important role in many of our investigations. Because vector bundles are not always treated in beginning manifolds courses, we include a fairly complete discussion of them in this chapter. The chapter ends with an application of these ideas to tensor bundles on manifolds, which are vector bundles constructed from tensor spaces associated with the tangent space at each point.

Much of the material included in this chapter should be familiar from your study of manifolds. It is included here as a review and to establish our notations and conventions for later use. If you need more detail on any topics mentioned here, consult [Boo86] or [Spi79, volume 1].

Tensors on a Vector Space

Let V be a finite-dimensional vector space (all our vector spaces and manifolds are assumed real). As usual, V^* denotes the dual space of V—the space of *covectors,* or real-valued linear functionals, on V—and we denote the natural pairing $V^* \times V \to \mathbf{R}$ by either of the notations

$$(\omega, X) \mapsto \langle \omega, X \rangle \quad \text{or} \quad (\omega, X) \mapsto \omega(X)$$

for $\omega \in V^*$, $X \in V$.

A *covariant k-tensor* on V is a multilinear map

$$F\colon \underbrace{V \times \cdots \times V}_{k \text{ copies}} \to \mathbf{R}.$$

Similarly, a *contravariant l-tensor* is a multilinear map

$$F\colon \underbrace{V^* \times \cdots \times V^*}_{l \text{ copies}} \to \mathbf{R}.$$

We often need to consider tensors of mixed types as well. A *tensor of type* $\binom{k}{l}$, also called a *k-covariant, l-contravariant tensor,* is a multilinear map

$$F\colon \underbrace{V^* \times \cdots \times V^*}_{l \text{ copies}} \times \underbrace{V \times \cdots \times V}_{k \text{ copies}} \to \mathbf{R}.$$

Actually, in many cases it is necessary to consider multilinear maps whose arguments consist of k vectors and l covectors, but not necessarily in the order implied by the definition above; such an object is still called a tensor of type $\binom{k}{l}$. For any given tensor, we will make it clear which arguments are vectors and which are covectors.

The space of all covariant k-tensors on V is denoted by $T^k(V)$, the space of contravariant l-tensors by $T_l(V)$, and the space of mixed $\binom{k}{l}$-tensors by $T^k_l(V)$. The *rank* of a tensor is the number of arguments (vectors and/or covectors) it takes.

There are obvious identifications $T^k_0(V) = T^k(V)$, $T^0_l(V) = T_l(V)$, $T^1(V) = V^*$, $T_1(V) = V^{**} = V$, and $T^0(V) = \mathbf{R}$. A less obvious, but extremely important, identification is $T^1_1(V) = \operatorname{End}(V)$, the space of linear endomorphisms of V (linear maps from V to itself). A more general version of this identification is expressed in the following lemma.

Lemma 2.1. *Let V be a finite-dimensional vector space. There is a natural (basis-independent) isomorphism between $T^k_{l+1}(V)$ and the space of multilinear maps*

$$\underbrace{V^* \times \cdots \times V^*}_{l} \times \underbrace{V \times \cdots \times V}_{k} \to V.$$

Exercise 2.1. Prove Lemma 2.1. [Hint: In the special case $k = 1$, $l = 0$, consider the map $\Phi\colon \operatorname{End}(V) \to T^1_1(V)$ by letting ΦA be the $\binom{1}{1}$-tensor defined by $\Phi A(\omega, X) = \omega(AX)$. The general case is similar.]

There is a natural product, called the *tensor product,* linking the various tensor spaces over V; if $F \in T^k_l(V)$ and $G \in T^p_q(V)$, the tensor $F \otimes G \in T^{k+p}_{l+q}(V)$ is defined by

$$\begin{aligned} F \otimes G(\omega^1, \ldots, \omega^{l+q}, X_1, \ldots, X_{k+p}) \\ = F(\omega^1, \ldots, \omega^l, X_1, \ldots, X_k)G(\omega^{l+1}, \ldots, \omega^{l+q}, X_{k+1}, \ldots, X_{k+p}). \end{aligned}$$

If $(E_1, \dots, E_n)$ is a basis for V, we let $(\varphi^1, \dots, \varphi^n)$ denote the corresponding dual basis for V^*, defined by $\varphi^i(E_j) = \delta^i_j$. A basis for $T^k_l(V)$ is given by the set of all tensors of the form

$$E_{j_1} \otimes \cdots \otimes E_{j_l} \otimes \varphi^{i_1} \otimes \cdots \otimes \varphi^{i_k}, \tag{2.1}$$

as the indices i_p, j_q range from 1 to n. These tensors act on basis elements by

$$E_{j_1} \otimes \cdots \otimes E_{j_l} \otimes \varphi^{i_1} \otimes \cdots \otimes \varphi^{i_k}(\varphi^{s_1}, \dots, \varphi^{s_l}, E_{r_1}, \dots, E_{r_k}) \\ = \delta^{s_1}_{j_1} \cdots \delta^{s_l}_{j_l} \delta^{i_1}_{r_1} \cdots \delta^{i_k}_{r_k}.$$

Any tensor $F \in T^k_l(V)$ can be written in terms of this basis as

$$F = F^{j_1 \dots j_l}_{i_1 \dots i_k} E_{j_1} \otimes \cdots \otimes E_{j_l} \otimes \varphi^{i_1} \otimes \cdots \otimes \varphi^{i_k}, \tag{2.2}$$

where

$$F^{j_1 \dots j_l}_{i_1 \dots i_k} = F(\varphi^{j_1}, \dots, \varphi^{j_l}, E_{i_1}, \dots, E_{i_k}).$$

In (2.2), and throughout this book, we use the *Einstein summation convention* for expressions with indices: if in any term the same index name appears twice, as both an upper and a lower index, that term is assumed to be summed over all possible values of that index (usually from 1 to the dimension of the space). We always choose our index positions so that vectors have lower indices and covectors have upper indices, while the *components* of vectors have upper indices and those of covectors have lower indices. This ensures that summations that make mathematical sense always obey the rule that each repeated index appears once up and once down in each term to be summed.

If the arguments of a mixed tensor F occur in a nonstandard order, then the horizontal as well as vertical positions of the indices are significant and reflect which arguments are vectors and which are covectors. For example, if B is a $\binom{2}{1}$-tensor whose first argument is a vector, second is a covector, and third is a vector, its components are written

$$B_i{}^j{}_k = B(E_i, \varphi^j, E_k). \tag{2.3}$$

We can use the result of Lemma 2.1 to define a natural operation called *trace* or *contraction,* which lowers the rank of a tensor by 2. In one special case, it is easy to describe: the operator $\operatorname{tr}\colon T^1_1(V) \to \mathbf{R}$ is just the trace of F when it is considered as an endomorphism of V. Since the trace of an endomorphism is basis-independent, this is well defined. More generally, we define $\operatorname{tr}\colon T^{k+1}_{l+1}(V) \to T^k_l(V)$ by letting $\operatorname{tr} F(\omega^1, \dots, \omega^l, V_1, \dots, V_k)$ be the trace of the endomorphism

$$F(\omega^1, \dots, \omega^l, \cdot, V_1, \dots, V_k, \cdot) \in T^1_1(V).$$

In terms of a basis, the components of $\operatorname{tr} F$ are

$$(\operatorname{tr} F)_{i_1\dots i_k}^{j_1\dots j_l} = F_{i_1\dots i_k m}^{j_1\dots j_l m}.$$

Even more generally, we can contract a given tensor on any pair of indices as long as one is contravariant and one is covariant. There is no general notation for this operation, so we just describe it in words each time it arises. For example, we can contract the tensor B with components given by (2.3) on its first and second indices to obtain a covariant 1-tensor A whose components are $A_k = B_i{}^i{}_k$.

Exercise 2.2. Show that the trace on any pair of indices is a well-defined linear map from $T_{l+1}^{k+1}(V)$ to $T_l^k(V)$.

A class of tensors that plays a special role in differential geometry is that of *alternating tensors:* those that change sign whenever two arguments are interchanged. We let $\Lambda^k(V)$ denote the space of covariant alternating k-tensors on V, also called *k-covectors* or *(exterior) k-forms*. There is a natural bilinear, associative product on forms called the *wedge product,* defined on 1-forms $\omega^1, \dots, \omega^k$ by setting

$$\omega^1 \wedge \dots \wedge \omega^k(X_1, \dots, X_k) = \det(\langle \omega^i, X_j \rangle),$$

and extending by linearity. (There is an alternative definition of the wedge product in common use, which amounts to multiplying our wedge product by a factor of $1/k!$. The choice of which definition to use is a matter of convention, though there are various reasons to justify each choice depending on the context. The definition we have chosen is most common in introductory differential geometry texts, and is used, for example, in [Boo86, Cha93, dC92, Spi79]. The other convention is used in [KN63] and is more common in complex differential geometry.)

Manifolds

Now we turn our attention to manifolds. Throughout this book, all our manifolds are assumed to be smooth, Hausdorff, and second countable; and *smooth* always means C^∞, or infinitely differentiable. As in most parts of differential geometry, the theory still works under weaker differentiability assumptions, but such considerations are usually relevant only when treating questions of hard analysis that are beyond our scope.

We write local coordinates on any open subset $U \subset M$ as $(x^1, \dots, x^n)$, (x^i), or x, depending on context. Although, formally speaking, coordinates constitute a map from U to $\mathbf{R}^n$, it is more common to use a coordinate chart to *identify* U with its image in $\mathbf{R}^n$, and to identify a point in U with its coordinate representation (x^i) in $\mathbf{R}^n$.

For any $p \in M$, the *tangent space* T_pM can be characterized either as the set of derivations of the algebra of germs at p of C^∞ functions on M (i.e., tangent vectors are "directional derivatives"), or as the set of equivalence classes of curves through p under a suitable equivalence relation (i.e., tangent vectors are "velocities"). Regardless of which characterization is taken as the definition, local coordinates (x^i) give a basis for T_pM consisting of the partial derivative operators $\partial/\partial x^i$. When there can be no confusion about which coordinates are meant, we usually abbreviate $\partial/\partial x^i$ by the notation ∂_i.

On a finite-dimensional vector space V with its standard smooth manifold structure, there is a natural (basis-independent) identification of each tangent space T_pV with V itself, obtained by identifying a vector $X \in V$ with the directional derivative

$$Xf = \frac{d}{dt}\bigg|_{t=0} f(p + tX).$$

In terms of the coordinates (x^i) induced on V by any basis, this is just the usual identification $(x^1, \ldots, x^n) \leftrightarrow x^i\partial_i$.

In this book, we always write coordinates with upper indices, as in (x^i). This has the consequence that the differentials dx^i of the coordinate functions are consistent with the convention that covectors have upper indices. Likewise, the coordinate vectors $\partial_i = \partial/\partial x^i$ have lower indices if we consider an upper index "in the denominator" to be the same as a lower index.

If $\widetilde{M}$ is a smooth manifold, a *submanifold* (or *immersed submanifold*) of $\widetilde{M}$ is a smooth manifold M together with an injective immersion $\iota\colon M \to \widetilde{M}$. Identifying M with its image $\iota(M) \subset \widetilde{M}$, we can consider M as a subset of $\widetilde{M}$, although in general the topology and smooth structure of M may have little to do with those of $\widetilde{M}$ and have to be considered as extra data. The most important type of submanifold is that in which the inclusion map ι is an *embedding,* which means that it is a homeomorphism onto its image with the subspace topology. In that case, M is called an *embedded submanifold* or a *regular submanifold.*

Suppose M is an embedded n-dimensional submanifold of an m-dimensional manifold $\widetilde{M}$. For every point $p \in M$, there exist *slice coordinates* $(x^1, \ldots, x^m)$ on a neighborhood $\widetilde{\mathcal{U}}$ of p in $\widetilde{M}$ such that $\widetilde{\mathcal{U}} \cap M$ is given by $\{x : x^{n+1} = \cdots = x^m = 0\}$, and $(x^1, \ldots, x^n)$ form local coordinates for M (Figure 2.1). At each $q \in \widetilde{\mathcal{U}} \cap M$, T_qM can be naturally identified as the subspace of $T_q\widetilde{M}$ spanned by the vectors $(\partial_1, \ldots, \partial_n)$.

Exercise 2.3. Suppose $M \subset \widetilde{M}$ is an embedded submanifold.

(a) If f is any smooth function on M, show that f can be extended to a smooth function on $\widetilde{M}$ whose restriction to M is f. [Hint: Extend f locally in slice coordinates by letting it be independent of $(x^{n+1}, \ldots, x^m)$, and patch together using a partition of unity.]

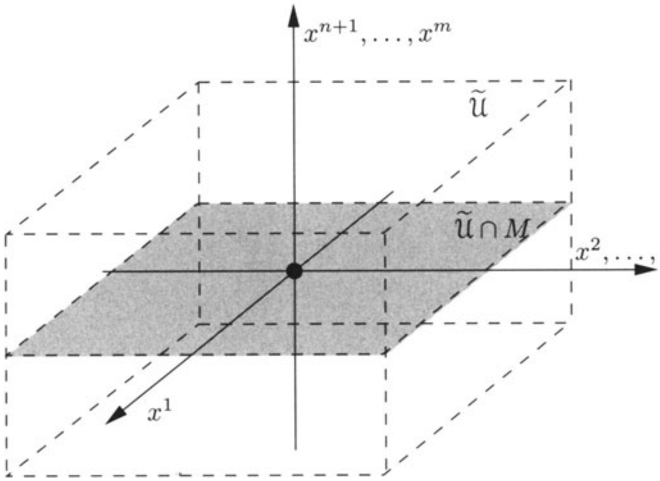


FIGURE 2.1. Slice coordinates.

(b) Show that any vector field on M can be extended to a vector field on $\widetilde{M}$.

(c) If $\widetilde{X}$ is a vector field on $\widetilde{M}$, show that $\widetilde{X}$ is tangent to M at points of M if and only if $\widetilde{X}f = 0$ whenever $f \in C^\infty(\widetilde{M})$ is a function that vanishes on M.

Vector Bundles

When we glue together the tangent spaces at all points on a manifold M, we get a set that can be thought of both as a union of vector spaces and as a manifold in its own right. This kind of structure is so common in differential geometry that it has a name.

A (*smooth*) *k-dimensional vector bundle* is a pair of smooth manifolds E (the *total space*) and M (the *base*), together with a surjective map $\pi\colon E \to M$ (the *projection*), satisfying the following conditions:

(a) Each set $E_p := \pi^{-1}(p)$ (called the *fiber* of E over p) is endowed with the structure of a vector space.

(b) For each $p \in M$, there exists a neighborhood U of p and a diffeomorphism $\varphi\colon \pi^{-1}(U) \to U \times \mathbf{R}^k$ (Figure 2.2), called a *local trivialization*

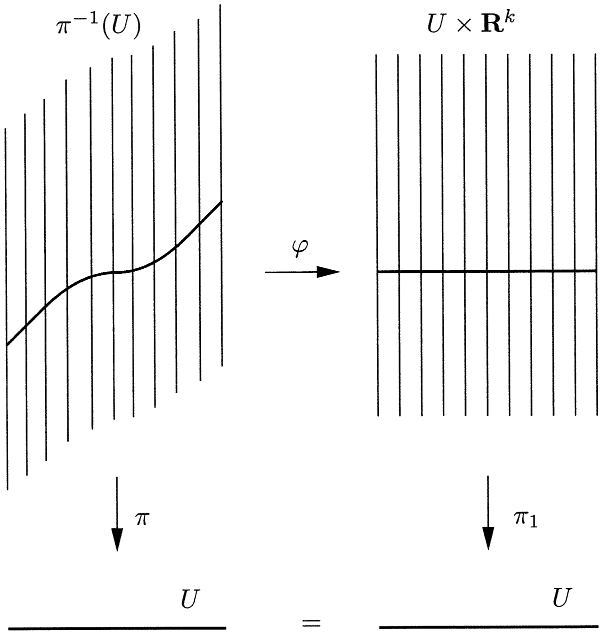


FIGURE 2.2. A local trivialization.

of E, such that the following diagram commutes:

$$\begin{array}{ccc} \pi^{-1}(U) & \xrightarrow{\varphi} & U \times \mathbf{R}^k \\ {\scriptstyle \pi}\downarrow & & \downarrow{\scriptstyle \pi_1} \\ U & = & U \end{array}$$

(where π_1 is the projection onto the first factor).

(c) The restriction of φ to each fiber, $\varphi \colon E_p \to \{p\} \times \mathbf{R}^k$, is a linear isomorphism.

Whether or not you have encountered the formal definition of vector bundles, you have certainly seen at least two examples: the *tangent bundle* TM of a smooth manifold M, which is just the disjoint union of the tangent spaces T_pM for all $p \in M$, and the *cotangent bundle* T^*M, which is the disjoint union of the cotangent spaces $T_p^*M = (T_pM)^*$. Another example that is relatively easy to visualize (and which we formally define in Chapter 8) is the *normal bundle* to a submanifold $M \subset \mathbf{R}^n$, whose fiber at each point is the normal space N_pM, the orthogonal complement of T_pM in $\mathbf{R}^n$.

It frequently happens that we are given a collection of vector spaces, one for each point in a manifold, that we would like to "glue together" to form a

vector bundle. For example, this is how the tangent and cotangent bundles are defined. There is a shortcut for showing that such a collection forms a vector bundle without first constructing a smooth manifold structure on the total space. As the next lemma shows, all we need to do is to exhibit the maps that we wish to consider as local trivializations and check that they overlap correctly.

Lemma 2.2. *Let M be a smooth manifold, E a set, and $\pi\colon E \to M$ a surjective map. Suppose we are given an open covering $\{U_\alpha\}$ of M together with bijective maps $\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times \mathbf{R}^k$ satisfying $\pi_1 \circ \varphi_\alpha = \pi$, such that whenever $U_\alpha \cap U_\beta \neq \emptyset$, the composite map*

$$\varphi_\alpha \circ \varphi_\beta^{-1}\colon U_\alpha \cap U_\beta \times \mathbf{R}^k \to U_\alpha \cap U_\beta \times \mathbf{R}^k$$

is of the form

$$\varphi_\alpha \circ \varphi_\beta^{-1}(p, V) = (p, \tau(p)V) \tag{2.4}$$

for some smooth map $\tau\colon U_\alpha \cap U_\beta \to GL(k, \mathbf{R})$. Then E has a unique structure as a smooth k-dimensional vector bundle over M for which the maps φ_α are local trivializations.

Proof. For each $p \in M$, let $E_p = \pi^{-1}(p)$. If $p \in U_\alpha$, observe that the map $(\varphi_\alpha)_p\colon E_p \to \{p\} \times \mathbf{R}^k$ obtained by restricting φ_α is a bijection. We can define a vector space structure on E_p by declaring this map to be a linear isomorphism. This structure is well defined, since for any other set U_β containing p, (2.4) guarantees that $(\varphi_\alpha)_p \circ (\varphi_\beta)_p^{-1} = \tau(p)$ is an isomorphism.

Shrinking the sets U_α and taking more of them if necessary, we may assume each of them is diffeomorphic to some open set $\widetilde{U}_\alpha \subset \mathbf{R}^n$. Following φ_α with such a diffeomorphism, we get a bijection $\pi^{-1}(U_\alpha) \to \widetilde{U}_\alpha \times \mathbf{R}^k$, which we can use as a coordinate chart for E. Because (2.4) shows that the φ_αs overlap smoothly, these charts determine a locally Euclidean topology and a smooth manifold structure on E. It is immediate that each map φ_α is a diffeomorphism with respect to this smooth structure, and the rest of the conditions for a vector bundle follow automatically. □

The smooth $GL(k, \mathbf{R})$-valued maps τ of the preceding lemma are called *transition functions* for E.

As an illustration, we show how to apply this construction to the tangent bundle. Given a coordinate chart $(U, (x^i))$ for M, any tangent vector $V \in T_xM$ at a point $x \in U$ can be expressed in terms of the coordinate basis as $V = v^i\partial/\partial x^i$ for some n-tuple $v = (v^1, \ldots, v^n)$. Define a bijection $\varphi\colon \pi^{-1}(U) \to U \times \mathbf{R}^n$ by sending $V \in T_xM$ to (x, v). Where two coordinate charts (x^i) and $(\tilde{x}^i)$ overlap, the respective coordinate basis vectors are related by

$$\frac{\partial}{\partial x^i} = \frac{\partial \tilde{x}^j}{\partial x^i}\frac{\partial}{\partial \tilde{x}^j},$$

and therefore the same vector V is represented by

$$V = \tilde{v}^j \frac{\partial}{\partial \tilde{x}^j} = v^i \frac{\partial}{\partial x^i} = v^i \frac{\partial \tilde{x}^j}{\partial x^i} \frac{\partial}{\partial \tilde{x}^j}.$$

This means that $\tilde{v}^j = v^i \partial \tilde{x}^j / \partial x^i$, so the corresponding local trivializations φ and $\widetilde{\varphi}$ are related by

$$\widetilde{\varphi} \circ \varphi^{-1}(x, v) = \widetilde{\varphi}(V) = (x, \tilde{v}) = (x, \tau(x)v),$$

where $\tau(x)$ is the $GL(n, \mathbf{R})$-valued function $\partial \tilde{x}^j / \partial x^i$. It is now immediate from Lemma 2.2 that these are the local trivializations for a vector bundle structure on TM.

It is useful to note that this construction actually gives explicit coordinates (x^i, v^i) on $\pi^{-1}(U)$, which we will refer to as *standard coordinates* for the tangent bundle.

If $\pi \colon E \to M$ is a vector bundle over M, a *section* of E is a map $F \colon M \to E$ such that $\pi \circ F = Id_M$, or, equivalently, $F(p) \in E_p$ for all p. It is said to be a *smooth section* if it is smooth as a map between manifolds. The next lemma gives another criterion for smoothness that is more easily verified in practice.

Lemma 2.3. *Let $F \colon M \to E$ be a section of a vector bundle. F is smooth if and only if the components $F^{j_1 \ldots j_l}_{i_1 \ldots i_k}$ of F in terms of any smooth local frame $\{E_i\}$ on an open set $U \in M$ depend smoothly on $p \in U$.*

Exercise 2.4. Prove Lemma 2.3.

The set of smooth sections of a vector bundle is an infinite-dimensional vector space under pointwise addition and multiplication by constants, whose zero element is the *zero section* ζ defined by $\zeta_p = 0 \in E_p$ for all $p \in M$. In this book, we use the script letter corresponding to the name of a vector bundle to denote its space of sections. Thus, for example, the space of smooth sections of TM is denoted $\mathcal{T}(M)$; it is the space of smooth vector fields on M. (Many books use the notation $\mathcal{X}(M)$ for this space, but our notation is more systematic, and seems to be becoming more common.)

Tensor Bundles and Tensor Fields

On a manifold M, we can perform the same linear-algebraic constructions on each tangent space T_pM that we perform on any vector space, yielding tensors at p. For example, a $\binom{k}{l}$-tensor at $p \in M$ is just an element of $T^k_l(T_pM)$. We define the *bundle of $\binom{k}{l}$-tensors* on M as

$$T^k_l M := \coprod_{p \in M} T^k_l(T_pM),$$

where $\coprod$ denotes the disjoint union. Similarly, the *bundle of k-forms* is

$$\Lambda^k M := \coprod_{p\in M} \Lambda^k(T_pM).$$

There are the usual identifications $T_1M = TM$ and $T^1M = \Lambda^1M = T^*M$.

To see that each of these tensor bundles is a vector bundle, define the projection $\pi\colon T^k_lM \to M$ to be the map that simply sends $F \in T^k_l(T_pM)$ to p. If (x^i) are any local coordinates on $U \subset M$, and $p \in U$, the coordinate vectors $\{\partial_i\}$ form a basis for T_pM whose dual basis is $\{dx^i\}$. Any tensor $F \in T^k_l(T_pM)$ can be expressed in terms of this basis as

$$F = F^{j_1\dots j_l}_{i_1\dots i_k}\,\partial_{j_1} \otimes \cdots \otimes \partial_{j_l} \otimes dx^{i_1} \otimes \cdots \otimes dx^{i_k}.$$

Exercise 2.5. For any coordinate chart $(U,(x^i))$ on M, define a map φ from $\pi^{-1}(U) \subset T^k_lM$ to $U \times \mathbf{R}^{n^{k+l}}$ by sending a tensor $F \in T^k_l(T_xM)$ to $(x,(F^{j_1\dots j_l}_{i_1\dots i_k})) \in U \times \mathbf{R}^{n^{k+l}}$. Show that T^k_lM can be made into a smooth vector bundle in a unique way so that all such maps φ are local trivializations.

A *tensor field* on M is a smooth section of some tensor bundle T^k_lM, and a *differential k-form* is a smooth section of Λ^kM. To avoid confusion between the point $p \in M$ at which a tensor field is evaluated and the vectors and covectors to which it is applied, we usually write the value of a tensor field F at $p \in M$ as $F_p \in T^k_l(T_pM)$, or, if it is clearer (for example if F itself has one or more subscripts), as $F|_p$. The space of $\binom{k}{l}$-tensor fields is denoted by $\mathcal{T}^k_l(M)$, and the space of covariant k-tensor fields (smooth sections of T^kM) by $\mathcal{T}^k(M)$. In particular, $\mathcal{T}^1(M)$ is the space of 1-forms. We follow the common practice of denoting the space of smooth real-valued functions on M (i.e., smooth sections of T^0M) by $C^\infty(M)$.

Let $(E_1,\dots,E_n)$ be any *local frame* for TM, that is, n smooth vector fields defined on some open set U such that $(E_1|_p,\dots,E_n|_p)$ form a basis for T_pM at each point $p \in U$. Associated with such a frame is the *dual coframe,* which we denote $(\varphi^1,\dots,\varphi^n)$; these are smooth 1-forms satisfying $\varphi^i(E_j) = \delta^i_j$. In terms of any local frame, a $\binom{k}{l}$-tensor field F can be written in the form (2.2), where now the components $F^{j_1\dots j_l}_{i_1\dots i_k}$ are to be interpreted as functions on U. In particular, in terms of a coordinate frame $\{\partial_i\}$ and its dual coframe $\{dx^i\}$, F has the coordinate expression

$$F_p = F^{j_1\dots j_l}_{i_1\dots i_k}(p)\,\partial_{j_1} \otimes \cdots \otimes \partial_{j_l} \otimes dx^{i_1} \otimes \cdots \otimes dx^{i_k}. \tag{2.5}$$

Exercise 2.6. Let $F\colon M \to T^k_lM$ be a section. Show that F is a smooth tensor field if and only if whenever $\{X_i\}$ are smooth vector fields and $\{\omega^j\}$ are smooth 1-forms defined on an open set $U \subset M$, the function $F(\omega^1,\dots,\omega^l,X_1,\dots,X_k)$ on U, defined by

$$F(\omega^1,\dots,\omega^l,X_1,\dots,X_k)(p) = F_p(\omega^1_p,\dots,\omega^l_p,X_1|_p,\dots,X_k|_p),$$

is smooth.

An important property of tensor fields is that they are multilinear over the space of smooth functions. Given a tensor field $F \in \mathcal{T}^k_l(M)$, vector fields $X_i \in \mathcal{T}(M)$, and 1-forms $\omega^j \in \mathcal{T}^1(M)$, Exercise 2.6 shows that the function $F(X_1, \ldots, X_k, \omega^1, \ldots, \omega^l)$ is smooth, and thus F induces a map

$$F\colon \mathcal{T}^1(M) \times \cdots \times \mathcal{T}^1(M) \times \mathcal{T}(M) \times \cdots \times \mathcal{T}(M) \to C^\infty(M).$$

It is easy to check that this map is *multilinear over* $C^\infty(M)$, that is, for any functions $f, g \in C^\infty(M)$ and any smooth vector or covector fields α, β,

$$F(\ldots, f\alpha + g\beta, \ldots) = fF(\ldots, \alpha, \ldots) + gF(\ldots, \beta, \ldots).$$

Even more important is the converse: as the next lemma shows, any such map that is multilinear over $C^\infty(M)$ defines a tensor field.

Lemma 2.4. (Tensor Characterization Lemma) *A map*

$$\tau\colon \mathcal{T}^1(M) \times \cdots \times \mathcal{T}^1(M) \times \mathcal{T}(M) \times \cdots \times \mathcal{T}(M) \to C^\infty(M)$$

is induced by a $\binom{k}{l}$*-tensor field as above if and only if it is multilinear over* $C^\infty(M)$*. Similarly, a map*

$$\tau\colon \mathcal{T}^1(M) \times \cdots \times \mathcal{T}^1(M) \times \mathcal{T}(M) \times \cdots \times \mathcal{T}(M) \to \mathcal{T}(M)$$

is induced by a $\binom{k}{l+1}$*-tensor field as in Lemma 2.1 if and only if it is multilinear over* $C^\infty(M)$.

Exercise 2.7. Prove Lemma 2.4.

3

Definitions and Examples of Riemannian Metrics

In this chapter we officially define Riemannian metrics and construct some of the elementary objects associated with them. At the end of the chapter, we introduce three classes of highly symmetric "model" Riemannian manifolds—Euclidean spaces, spheres, and hyperbolic spaces—to which we will return repeatedly as our understanding deepens and our tools become more sophisticated.

Riemannian Metrics

Definitions

A *Riemannian metric* on a smooth manifold M is a 2-tensor field $g \in \mathcal{T}^2(M)$ that is *symmetric* (i.e., $g(X,Y) = g(Y,X)$) and *positive definite* (i.e., $g(X,X) > 0$ if $X \neq 0$). A Riemannian metric thus determines an inner product on each tangent space T_pM, which is typically written $\langle X, Y \rangle := g(X,Y)$ for $X, Y \in T_pM$. A manifold together with a given Riemannian metric is called a *Riemannian manifold.* We often use the word "metric" to refer to a Riemannian metric when there is no chance of confusion.

Exercise 3.1. Using a partition of unity, prove that every manifold can be given a Riemannian metric.

Just as in Euclidean geometry, if p is a point in a Riemannian manifold (M, g), we define the *length* or *norm* of any tangent vector $X \in T_pM$ to be $|X| := \langle X, X \rangle^{1/2}$. Unless we specify otherwise, we define the *angle* between

two nonzero vectors $X, Y \in T_pM$ to be the unique $\theta \in [0, \pi]$ satisfying $\cos\theta = \langle X, Y\rangle/(|X|\,|Y|)$. (Later, we will further refine the notion of angle in special cases to allow more general values of θ.) We say that X and Y are *orthogonal* if their angle is $\pi/2$, or equivalently if $\langle X, Y\rangle = 0$. Vectors $E_1, \ldots, E_k$ are called *orthonormal* if they are of length 1 and pairwise orthogonal, or equivalently if $\langle E_i, E_j\rangle = \delta_{ij}$.

If (M, g) and $(\widetilde{M}, \tilde{g})$ are Riemannian manifolds, a diffeomorphism φ from M to $\widetilde{M}$ is called an *isometry* if $\varphi^*\tilde{g} = g$. We say (M, g) and $(\widetilde{M}, \tilde{g})$ are *isometric* if there exists an isometry between them. It is easy to verify that being isometric is an equivalence relation on the class of Riemannian manifolds. Riemannian geometry is concerned primarily with properties that are preserved by isometries.

An isometry $\varphi\colon (M, g) \to (M, g)$ is called an *isometry of M*. A composition of isometries and the inverse of an isometry are again isometries, so the set of isometries of M is a group, called the *isometry group* of M; it is denoted $\mathcal{I}(M)$. (It can be shown that the isometry group is always a finite-dimensional Lie group acting smoothly on M; see, for example, [Kob72, Theorem II.1.2].)

If $(E_1, \ldots, E_n)$ is any local frame for TM, and $(\varphi^1, \ldots, \varphi^n)$ is its dual coframe, a Riemannian metric can be written locally as

$$g = g_{ij}\varphi^i \otimes \varphi^j.$$

The coefficient matrix, defined by $g_{ij} = \langle E_i, E_j\rangle$, is symmetric in i and j and depends smoothly on $p \in M$. In particular, in a coordinate frame, g has the form

$$g = g_{ij}dx^i \otimes dx^j. \tag{3.1}$$

The notation can be shortened by introducing the *symmetric product* of two 1-forms ω and η, denoted by juxtaposition with no product symbol:

$$\omega\eta := \tfrac{1}{2}(\omega \otimes \eta + \eta \otimes \omega).$$

Because of the symmetry of g_{ij}, (3.1) is equivalent to

$$g = g_{ij}dx^i dx^j.$$

Exercise 3.2. Let p be any point in a Riemannian n-manifold (M, g). Show that there is a *local orthonormal frame* near p—that is, a local frame $E_1, \ldots, E_n$ defined in a neighborhood of p that forms an orthonormal basis for the tangent space at each point. [Hint: Use the Gram–Schmidt algorithm. *Warning:* A common mistake made by novices is to assume that one can find coordinates near p such that the *coordinate* vector fields ∂_i are orthonormal. Your solution to this exercise *does not* show this. In fact, as we will see in Chapter 7, this is possible only when the metric is flat, i.e., locally isometric to the Euclidean metric.]

Examples

One obvious example of a Riemannian manifold is $\mathbf{R}^n$ with its *Euclidean metric* $\bar{g}$, which is just the usual inner product on each tangent space $T_x\mathbf{R}^n$ under the natural identification $T_x\mathbf{R}^n = \mathbf{R}^n$. In standard coordinates, this can be written in several ways:

$$\bar{g} = \sum_i dx^i dx^i = \sum_i (dx^i)^2 = \delta_{ij} dx^i dx^j. \tag{3.2}$$

The matrix of $\bar{g}$ in these coordinates is thus $\bar{g}_{ij} = \delta_{ij}$.

Many other examples of Riemannian metrics arise naturally as submanifolds, products, and quotients of Riemannian manifolds. We begin with submanifolds. Suppose $(\widetilde{M}, \tilde{g})$ is a Riemannian manifold, and $\iota\colon M \hookrightarrow \widetilde{M}$ is an (immersed) submanifold of $\widetilde{M}$. The *induced metric* on M is the 2-tensor $g = \iota^*\tilde{g}$, which is just the restriction of $\tilde{g}$ to vectors tangent to M. Because the restriction of an inner product is itself an inner product, this obviously defines a Riemannian metric on M. For example, the standard metric on the sphere $\mathbf{S}^n \subset \mathbf{R}^{n+1}$ is obtained in this way; we study it in much more detail later in this chapter.

Computations on a submanifold are usually most conveniently carried out in terms of a *local parametrization:* this is an embedding of an open subset $\mathcal{U} \subset \mathbf{R}^n$ into $\widetilde{M}$, whose image is an open subset of M. For example, if $X\colon \mathcal{U} \to \mathbf{R}^m$ is a parametrization of a submanifold $M \subset \mathbf{R}^m$ with the induced metric, the induced metric in standard coordinates $(u^1, \dots, u^n)$ on $\mathcal{U}$ is just

$$g = \sum_{i=1}^m (dX^i)^2 = \sum_{i=1}^m \left(\frac{\partial X^i}{\partial u^j} du^j \right)^2 .$$

Exercise 3.3. Let $\gamma(t) = (a(t), b(t))$, $t \in I$ (an open interval), be a smooth injective curve in the xz-plane, and suppose $a(t) > 0$ and $\dot{\gamma}(t) \neq 0$ for all $t \in I$. Let $M \subset \mathbf{R}^3$ be the surface of revolution obtained by revolving the image of γ about the z-axis (Figure 3.1).

(a) Show that M is an immersed submanifold of $\mathbf{R}^3$, and is embedded if γ is an embedding.

(b) Show that the map $\varphi(\theta, t) = (a(t)\cos\theta, a(t)\sin\theta, b(t))$ from $\mathbf{R} \times I$ to $\mathbf{R}^3$ is a local parametrization of M in a neighborhood of any point.

(c) Compute the expression for the induced metric on M in (θ, t) coordinates.

(d) Specialize this computation to the case of the doughnut-shaped torus of revolution given by $(a(t), b(t)) = (2 + \cos t, \sin t)$.

Exercise 3.4. The *n-torus* is the manifold $\mathbf{T}^n := S^1 \times \cdots \times S^1$, considered as the subset of $\mathbf{R}^{2n}$ defined by $(x^1)^2 + (x^2)^2 = \cdots = (x^{2n-1})^2 + (x^{2n})^2 = 1$. Show that $X(u^1, \dots, u^n) = (\cos u^1, \sin u^1, \dots, \cos u^n, \sin u^n)$ gives local

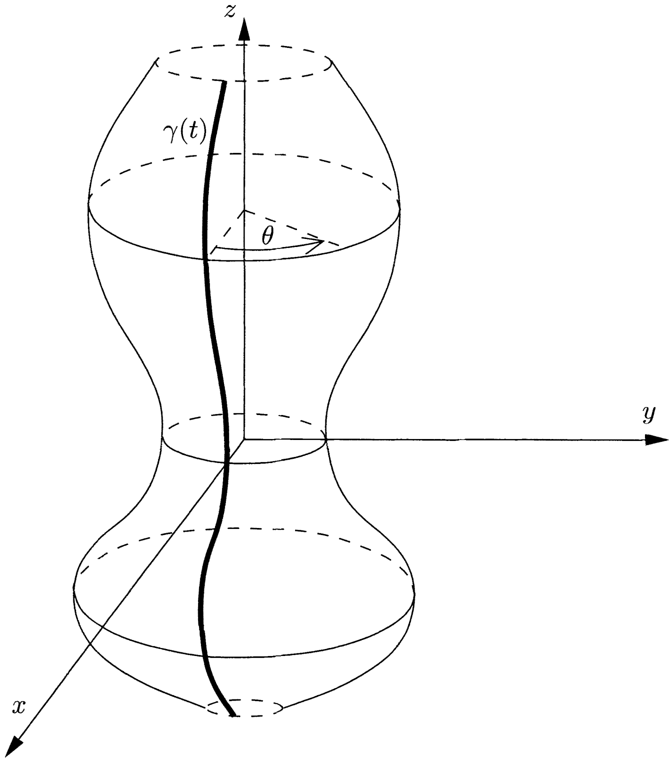


FIGURE 3.1. A surface of revolution.

parametrizations of $\mathbf{T}^n$ when restricted to suitable domains, and that the induced metric is equal to the Euclidean metric in (u^i) coordinates.

Next we consider products. If (M_1, g_1) and (M_2, g_2) are Riemannian manifolds, the product $M_1 \times M_2$ has a natural Riemannian metric $g = g_1 \oplus g_2$, called the *product metric,* defined by

$$g(X_1 + X_2, Y_1 + Y_2) = g_1(X_1, Y_1) + g_2(X_2, Y_2), \tag{3.3}$$

where $X_i, Y_i \in T_{p_i}M_i$ under the natural identification $T_{(p_1,p_2)}M_1 \times M_2 = T_{p_1}M_1 \oplus T_{p_2}M_2$.

Any local coordinates $(x^1, \ldots, x^n)$ for M_1 and $(x^{n+1}, \ldots, x^{n+m})$ for M_2 give coordinates $(x^1, \ldots, x^{n+m})$ for $M_1 \times M_2$. In terms of these coordinates, the product metric has the local expression $g = g_{ij}dx^i dx^j$, where (g_{ij}) is the block diagonal matrix

$$(g_{ij}) = \begin{pmatrix} (g_1)_{ij} & 0 \\ 0 & (g_2)_{ij} \end{pmatrix}.$$

Exercise 3.5. Show that the induced metric on $\mathbf{T}^n$ described in Exercise 3.4 is the product metric obtained from the usual induced metric on $\mathbf{S}^1 \subset \mathbf{R}^2$.

Our last class of examples is obtained from covering spaces. Suppose $\pi\colon \widetilde{M} \to M$ is a smooth covering map. A *covering transformation* (or *deck transformation*) is a smooth map $\varphi\colon \widetilde{M} \to \widetilde{M}$ such that $\pi \circ \varphi = \pi$. If g is a Riemannian metric on M, then $\tilde{g} := \pi^* g$ is a Riemannian metric on $\widetilde{M}$ that is invariant under all covering transformations. In this case $\tilde{g}$ is called the *covering metric,* and π is called a *Riemannian covering.*

The following exercise shows the converse: Any metric on $\widetilde{M}$ that is invariant under all covering transformations descends to M.

Exercise 3.6. If $\pi\colon \widetilde{M} \to M$ is a smooth covering map, and $\tilde{g}$ is any metric on $\widetilde{M}$ that is invariant under all covering transformations, show that there is a unique metric g on M such that $\tilde{g} = \pi^* g$.

Exercise 3.7. Let $\mathbf{T}^n \subset \mathbf{R}^{2n}$ denote the n-torus. Show that the map $X\colon \mathbf{R}^n \to \mathbf{T}^n$ of Exercise 3.4 is a Riemannian covering.

Later in this chapter, we will undertake a much more detailed study of three important classes of examples of Riemannian metrics, the "model spaces" of Riemannian geometry. Other examples, such as metrics on Lie groups and on complex projective spaces, are introduced in the problems at the end of the chapter.

Elementary Constructions Associated with Riemannian Metrics

Raising and Lowering Indices

One elementary but important property of Riemannian metrics is that they allow us to convert vectors to covectors and vice versa. Given a metric g on M, define a map called *flat* from TM to T^*M by sending a vector X to the covector $X^\flat$ defined by

$$X^\flat(Y) := g(X, Y).$$

In coordinates,

$$X^\flat = g\left(X^i \partial_i, \cdot\right) = g_{ij} X^i dx^j.$$

It is standard practice to write $X^\flat$ in coordinates as $X^\flat = X_j dx^j$, where

$$X_j := g_{ij} X^i.$$

One says that $X^\flat$ is obtained from X by *lowering an index.* (This is why the operation is designated by the musical notation $\flat$ = "flat.")

The matrix of flat in terms of a coordinate basis is therefore the matrix of g itself. Since the matrix of g is invertible, so is the flat operator; we denote its inverse by (what else?) $\omega \mapsto \omega^\#$, called *sharp*. In coordinates, $\omega^\#$ has components

$$\omega^i := g^{ij}\omega_j,$$

where, by definition, g^{ij} are the components of the inverse matrix $(g_{ij})^{-1}$. One says $\omega^\#$ is obtained by *raising an index.*

Probably the most important application of the sharp operator is to extend the classical gradient operator to Riemannian manifolds. If f is a smooth, real-valued function on a Riemannian manifold (M, g), the *gradient* of f is the vector field $\operatorname{grad} f := df^\#$ obtained from df by raising an index. Looking through the definitions, we see that $\operatorname{grad} f$ is characterized by the fact that

$$df(Y) = \langle \operatorname{grad} f, Y\rangle \qquad \text{for all } Y \in TM,$$

and has the coordinate expression

$$\operatorname{grad} f = g^{ij}\partial_i f \partial_j.$$

The flat and sharp operators can be applied to tensors of any rank, in any index position, to convert tensors from covariant to contravariant or vice versa. For example, if B is again the 3-tensor with components given by (2.3), we can lower its middle index to obtain a covariant 3-tensor $B^\flat$ with components

$$B_{ijk} := g_{jl}B_i{}^l{}_k.$$

In coordinate-free notation, this is just

$$B^\flat(X, Y, Z) := B(X, Y^\flat, Z).$$

(Of course, if a tensor has more than one upper index, the flat notation doesn't tell us which one to lower. In such cases, we have to explain in words what is meant.)

Another important application of the flat and sharp operators is to extend the trace operator introduced in Chapter 2 to covariant tensors. We consider only symmetric 2-tensors here, but it is easy to extend these results to more general tensors.

If h is a symmetric 2-tensor on a Riemannian manifold, then $h^\#$ is a $\binom{1}{1}$-tensor and therefore $\operatorname{tr} h^\#$ is defined. We define the *trace of h with respect to g* as

$$\operatorname{tr}_g h := \operatorname{tr} h^\#.$$

(Because h is symmetric, it doesn't matter which index is raised.) In terms of a basis, this is

$$\operatorname{tr}_g h = h_i{}^i = g^{ij} h_{ij}.$$

In particular, in an orthonormal basis this is the ordinary trace of a matrix.

Inner Products of Tensors

A metric is by definition an inner product on tangent vectors. As the following lemma shows, it determines an inner product (and hence a norm) on all tensor bundles as well. First a bit of terminology: If $E \to M$ is a vector bundle, a *fiber metric* on E is an inner product on each fiber E_p that varies smoothly, in the sense that for any (local) smooth sections σ, τ of E, the inner product $\langle \sigma, \tau \rangle$ is a smooth function.

Lemma 3.1. *Let g be a Riemannian metric on a manifold M. There is a unique fiber metric on each tensor bundle $T_l^k M$ with the property that if $(E_1, \ldots, E_n)$ is an orthonormal basis for T_pM and $(\varphi^1, \ldots, \varphi^n)$ is the corresponding dual basis, then the collection of tensors given by* (2.1) *forms an orthonormal basis for $T_l^k(T_pM)$.*

Exercise 3.8. Prove Lemma 3.1 by showing that in any local coordinate system, the required inner product is given by

$$\langle F, G \rangle = g^{i_1 r_1} \cdots g^{i_k r_k} g_{j_1 s_1} \cdots g_{j_l s_l} F_{i_1 \ldots i_k}^{j_1 \ldots j_l} G_{r_1 \ldots r_k}^{s_1 \ldots s_l}.$$

Show moreover that if ω, η are covariant 1-tensors, then

$$\langle \omega, \eta \rangle = \langle \omega^\#, \eta^\# \rangle.$$

The Volume Element and Integration

The final general construction we will study before looking at specific examples of metrics is the volume element.

Lemma 3.2. *On any oriented Riemannian n-manifold (M, g), there is a unique n-form dV satisfying the property that $dV(E_1, \ldots, E_n) = 1$ whenever $(E_1, \ldots, E_n)$ is an oriented orthonormal basis for some tangent space T_pM.*

This n-form dV (sometimes denoted dV_g for clarity) is called the *(Riemannian) volume element.*

Exercise 3.9. Prove Lemma 3.2, and show that the expression for dV with respect to any oriented local frame $\{E_i\}$ is

$$dV = \sqrt{\det(g_{ij})}\, \varphi^1 \wedge \cdots \wedge \varphi^n,$$

where $g_{ij} = \langle E_i, E_j \rangle$ are the coefficients of g and $\{\varphi^i\}$ is the dual coframe.

The significance of the Riemannian volume element is that it allows us to integrate functions, not just differential forms. If f is a smooth, compactly supported function on an oriented Riemannian n-manifold (M, g), then $f\,dV$ is a compactly supported n-form. Therefore the integral $\int_M f\,dV$ makes sense, and we define it to be the *integral of f over M.* Similarly, the *volume of M* is defined to be $\int_M dV = \int_M 1\,dV$.

Generalizations of Riemannian Metrics

There are other common ways of measuring "lengths" of tangent vectors on smooth manifolds. Let's digress briefly to mention three that play important roles in other branches of mathematics: pseudo-Riemannian metrics, sub-Riemannian metrics, and Finsler metrics. Each is defined by relaxing one of the requirements in the definition of Riemannian metric: a pseudo-Riemannian metric is obtained by relaxing the requirement that the metric be positive; a sub-Riemannian metric by relaxing the requirement that it be defined on the whole tangent space; and a Finsler metric by relaxing the requirement that it be quadratic on each tangent space.

Pseudo-Riemannian Metrics

A *pseudo-Riemannian metric* (occasionally also called a *semi-Riemannian metric*) on a smooth manifold M is a symmetric 2-tensor field g that is *nondegenerate* at each point $p \in M$. This means that the only vector orthogonal to everything is the zero vector. More formally, $g(X, Y) = 0$ for all $Y \in T_pM$ if and only if $X = 0$. If $g = g_{ij}\varphi^i\varphi^j$ in terms of a local coframe, nondegeneracy just means that the matrix g_{ij} is invertible. If g is Riemannian, nondegeneracy follows immediately from positive-definiteness, so every Riemannian metric is also a pseudo-Riemannian metric; but in general pseudo-Riemannian metrics need not be positive.

Given a pseudo-Riemannian metric g and a point $p \in M$, by a simple extension of the Gram–Schmidt algorithm one can construct a basis $(E_1, \ldots, E_n)$ for T_pM in which g has the expression

$$g = -(\varphi^1)^2 - \cdots - (\varphi^r)^2 + (\varphi^{r+1})^2 + \cdots + (\varphi^n)^2 \tag{3.4}$$

for some integer $0 \le r \le n$. This integer r, called the *index* of g, is equal to the maximum dimension of any subspace of T_pM on which g is negative definite. Therefore the index is independent of the choice of basis, a fact known classically as *Sylvester's law of inertia.*

By far the most important pseudo-Riemannian metrics (other than the Riemannian ones) are the *Lorentz metrics,* which are pseudo-Riemannian metrics of index 1. The most important example of a Lorentz metric is the

Minkowski metric; this is the Lorentz metric m on $\mathbf{R}^{n+1}$ that is written in terms of coordinates $(\xi^1, \ldots, \xi^n, \tau)$ as

$$m = (d\xi^1)^2 + \cdots + (d\xi^n)^2 - (d\tau)^2. \tag{3.5}$$

In the special case of $\mathbf{R}^4$, the Minkowski metric is the fundamental invariant of Einstein's special theory of relativity, which can be expressed succinctly by saying that in the absence of gravity, the laws of physics have the same form in any coordinate system in which the Minkowski metric has the expression (3.5). The differing physical characteristics of "space" (the ξ directions) and "time" (the τ direction) arise from the fact that they are subspaces on which g is positive definite and negative definite, respectively. The general theory of relativity includes gravitational effects by allowing the Lorentz metric to vary from point to point.

Many aspects of the theory of Riemannian metrics apply equally well to pseudo-Riemannian metrics. Although we do not treat pseudo-Riemannian geometry directly in this book, we will attempt to point out as we go along which aspects of the theory apply to pseudo-Riemannian metrics. As a rule of thumb, proofs that depend only on the invertibility of the metric tensor, such as existence and uniqueness of the Riemannian connection and geodesics, work fine in the pseudo-Riemannian setting, while proofs that use positivity in an essential way, such as those involving distance-minimizing properties of geodesics, do not.

For an introduction to the mathematical aspects of pseudo-Riemannian metrics, see the excellent book [O'N83]; a more physical treatment can be found in [HE73].

Sub-Riemannian Metrics

A *sub-Riemannian metric* (also sometimes known as a *singular Riemannian metric* or *Carnot–Carathéodory metric*) on a manifold M is a fiber metric on a smooth distribution $S \subset TM$ (i.e., a k-plane field or sub-bundle of TM). Since lengths make sense only for vectors in S, the only curves whose lengths can be measured are those whose tangent vectors lie everywhere in S. Therefore one usually imposes some condition on S that guarantees that any two nearby points can be connected by such a curve. This is, in a sense, the opposite of the Frobenius integrability condition, which would restrict every such curve to lie in a single leaf of a foliation.

Sub-Riemannian metrics arise naturally in the study of the abstract models of real submanifolds of complex space $\mathbf{C}^n$, called *CR manifolds.* (Here CR stands for "Cauchy–Riemann.") CR manifolds are real manifolds endowed with a distribution $S \subset TM$ whose fibers carry the structure of complex vector spaces (with an additional integrability condition that need not concern us here). In the model case of a submanifold $M \subset \mathbf{C}^n$, S is the set of vectors tangent to M that remain tangent after multiplication by $i = \sqrt{-1}$

in the ambient complex coordinates. If S is sufficiently far from being integrable, choosing a fiber metric on S results in a sub-Riemannian metric whose geometric properties closely reflect the complex-analytic properties of M as a subset of $\mathbf{C}^n$.

Another motivation for studying sub-Riemannian metrics arises from *control theory.* In this subject, one is given a manifold with a vector field depending on parameters called *controls,* with the goal being to vary the controls so as to obtain a solution curve with desired properties, often one that minimizes some function such as arc length. If the vector field is everywhere tangent to a distribution S on the manifold (for example, in the case of a robot arm whose motion is restricted by the orientations of its hinges), then the function can often be modeled as a sub-Riemannian metric and optimal solutions modeled as sub-Riemannian geodesics.

A useful introduction to the geometry of sub-Riemannian metrics is provided in the article [Str86].

Finsler Metrics

A *Finsler metric* on a manifold M is a continuous function $F\colon TM \to \mathbf{R}$, smooth on the complement of the zero section, that defines a norm on each tangent space T_pM. This means that $F(X) > 0$ for $X \neq 0$, $F(cX) = |c|F(X)$ for $c \in \mathbf{R}$, and $F(X+Y) \leq F(X) + F(Y)$. Again, the norm function associated with any Riemannian metric is a special case.

The inventor of Riemannian geometry himself, G. F. B. Riemann, clearly envisaged an important role in n-dimensional geometry for what we now call Finsler metrics; he restricted his investigations to the "Riemannian" case purely for simplicity (see [Spi79, volume 2]). However, only very recently have Finsler metrics begun to be studied seriously from a geometric point of view—see [Che96] for a survey of recent progress in the differential-geometric investigation of Finsler metrics.

The recent upsurge of interest in Finsler metrics has been motivated largely by the fact that two different Finsler metrics appear very naturally in the theory of several complex variables: at least for bounded strictly convex domains in $\mathbf{C}^n$, the *Kobayashi metric* and the *Carathéodory metric* are intrinsically defined, biholomorphically invariant Finsler metrics. Combining differential-geometric and complex-analytic methods has led to striking new insights into both the function theory and the geometry of such domains. We do not treat Finsler metrics further in this book, but you can consult one of the recent books on the subject (e.g. [AP94, JP93]) or the references cited in [Che96].

The Model Spaces of Riemannian Geometry

Before we delve into the general theory of Riemannian manifolds, let's give it some substance by introducing three classes of highly symmetric "model spaces" of Riemannian geometry—Euclidean space, spheres, and hyperbolic spaces. For much more information on the material covered in this section, see [Wol84].

Euclidean Space

The simplest and most important model Riemannian manifold is of course $\mathbf{R}^n$ itself, with the Euclidean metric $\bar{g}$ given by (3.2). More generally, if V is any n-dimensional vector space endowed with an inner product, we can set $g(X,Y) = \langle X, Y\rangle$ for any $X, Y \in T_pV = V$. Choosing an orthonormal basis $(E_1, \ldots, E_n)$ for V defines a map from $\mathbf{R}^n$ to V by sending $(x^1, \ldots, x^n)$ to x^iE_i; this is easily seen to be an isometry of (V, g) with $(\mathbf{R}^n, \bar{g})$.

Spheres

Our second model space is the sphere of radius R in $\mathbf{R}^{n+1}$, denoted $\mathbf{S}^n_R$, with the metric $\mathring{g}_R$ induced from the Euclidean metric on $\mathbf{R}^{n+1}$, which we call the *round metric of radius R.* (When $R = 1$, this is simply called the *round metric,* and we'll use the notations $\mathbf{S}^n$ and $\mathring{g}$.)

One of the first things one notices about the spheres is that they are highly symmetric. To describe the symmetries of the sphere, we introduce some standard terminology. Let M be a Riemannian manifold. First, M is a *homogeneous Riemannian manifold* if it admits a Lie group acting smoothly and transitively by isometries. Second, given a point $p \in M$, M is *isotropic at p* if there exists a Lie group G acting smoothly on M by isometries such that the isotropy subgroup $G_p \subset G$ (the subgroup of elements of G that fix p) acts transitively on the set of unit vectors in T_pM (where $g \in G_p$ acts on T_pM by $g_*\colon T_pM \to T_pM$). Clearly a homogeneous Riemannian manifold that is isotropic at one point is isotropic at every point; in that case, one says M is *homogeneous and isotropic.* A homogeneous Riemannian manifold looks geometrically the same at every point, while an isotropic one looks the same in every direction.

We can immediately write down a large group of isometries of $\mathbf{S}^n_R$ by observing that the linear action of the orthogonal group $O(n+1)$ on $\mathbf{R}^{n+1}$ preserves $\mathbf{S}^n_R$ and the Euclidean metric, so its restriction to $\mathbf{S}^n_R$ acts by isometries of the sphere. (Later we'll see in fact that this is the full isometry group, but we don't need that fact now.)

Proposition 3.3. *$O(n+1)$ acts transitively on orthonormal bases on $\mathbf{S}^n_R$. More precisely, given any two points $p, \tilde{p} \in \mathbf{S}^n_R$, and orthonormal bases $\{E_i\}$*

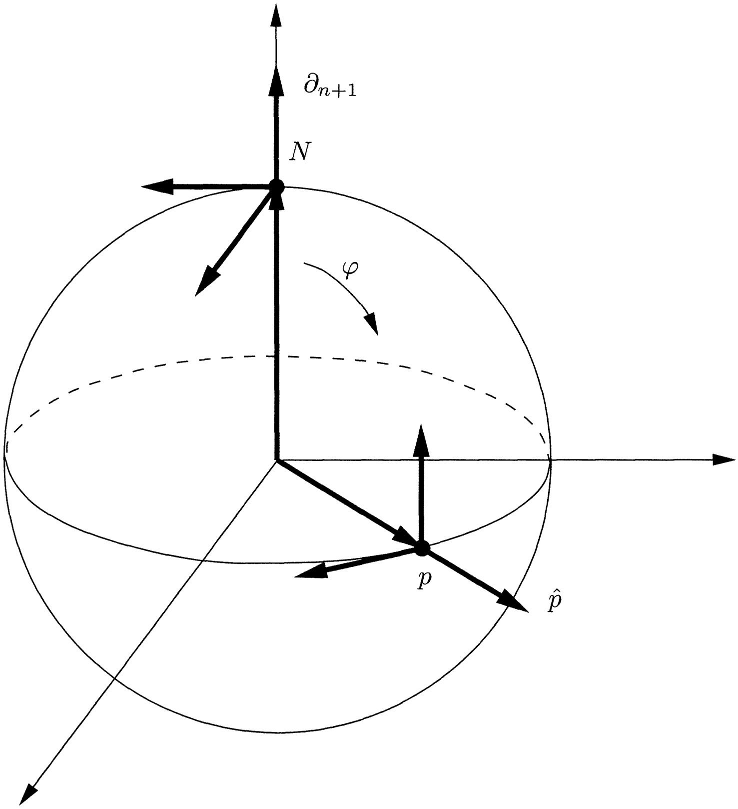


FIGURE 3.2. Transitivity of $O(n+1)$ on orthonormal bases.

for $T_p\mathbf{S}^n_R$ *and* $\{\tilde{E}_i\}$ *for* $T_{\tilde{p}}\mathbf{S}^n_R$*, there exists* $\varphi \in O(n+1)$ *such that* $\varphi(p) = \tilde{p}$ *and* $\varphi_* E_i = \tilde{E}_i$*. In particular,* $\mathbf{S}^n_R$ *is homogeneous and isotropic.*

Proof. It suffices to show that given any $p \in \mathbf{S}^n_R$ and any orthonormal basis $\{E_i\}$ for $T_p\mathbf{S}^n_R$, there is an orthogonal map that takes the "north pole" $N = (0, \dots, 0, R)$ to p and the standard basis $\{\partial_i\}$ for $T_N\mathbf{S}^n_R$ to $\{E_i\}$.

To do so, think of p as a vector of length R in $\mathbf{R}^{n+1}$, and let $\hat{p} = p/R$ denote the corresponding unit vector (Figure 3.2). Since the basis vectors $\{E_i\}$ are tangent to the sphere, they are orthogonal to $\hat{p}$, so $(E_1, \dots, E_n, \hat{p})$ is an orthonormal basis for $\mathbf{R}^{n+1}$. Let α be the matrix whose columns are these basis vectors. Then $\alpha \in O(n+1)$, and by elementary linear algebra α takes the standard basis vectors $(\partial_1, \dots, \partial_{n+1})$ to $(E_1, \dots, E_n, \hat{p})$. In particular, $\alpha(0, \dots, 0, R) = p$. Moreover, since α acts linearly on $\mathbf{R}^{n+1}$, its push-forward is represented in standard coordinates by the same matrix, so $\alpha_* \partial_i = E_i$ for $i = 1, \dots, n$, and α is the desired orthogonal map. □

Another important feature of the sphere—one that is much less evident than its symmetry—is that it is locally conformally equivalent to Euclidean space, in a sense that we now describe. Two metrics g_1 and g_2 on a manifold M are said to be *conformal* to each other if there is a positive function $f \in C^\infty(M)$ such that $g_2 = f g_1$. Two Riemannian manifolds (M, g) and $(\widetilde{M}, \tilde{g})$ are said to be *conformally equivalent* if there is a diffeomorphism $\varphi\colon M \to \widetilde{M}$ such that $\varphi^* \tilde{g}$ is conformal to g.

Exercise 3.10. (a) Show that two metrics are conformal if and only if they define the same angles but not necessarily the same lengths.

(b) Show that a diffeomorphism is a conformal equivalence if and only if it preserves angles.

A conformal equivalence between $\mathbf{R}^n$ and the sphere $\mathbf{S}^n_R \subset \mathbf{R}^{n+1}$ minus a point is provided by *stereographic projection* from the north pole. This is the map $\sigma\colon \mathbf{S}^n_R - \{N\} \to \mathbf{R}^n$ that sends a point $P \in \mathbf{S}^n_R - \{N\} \subset \mathbf{R}^{n+1}$, written $P = (\xi^1, \ldots, \xi^n, \tau)$, to $u \in \mathbf{R}^n$, where $U = (u^1, \ldots, u^n, 0)$ is the point where the line through N and P intersects the hyperplane $\{\tau = 0\}$ in $\mathbf{R}^{n+1}$ (Figure 3.3). Thus U is characterized by the fact that $\overrightarrow{NU} = \lambda \overrightarrow{NP}$ for some nonzero scalar λ. Writing $N = (0, R)$, $U = (u, 0)$, and $P = (\xi, \tau) \in \mathbf{R}^{n+1} = \mathbf{R}^n \times \mathbf{R}$, this leads to the system of equations

$$\begin{aligned} u^i &= \lambda \xi^i, \\ -R &= \lambda(\tau - R). \end{aligned} \tag{3.6}$$

Solving the second equation for λ and plugging it into the first equation, we get the formula for stereographic projection

$$\sigma(\xi, \tau) = u = \frac{R\xi}{R - \tau}. \tag{3.7}$$

Clearly σ is defined and smooth on all of $\mathbf{S}^n_R - \{N\}$. The easiest way to see that it is a diffeomorphism is to compute its inverse. Solving the two equations of (3.6) for τ and ξ^i gives

$$\xi^i = \frac{u^i}{\lambda}, \qquad \tau = R\frac{\lambda - 1}{\lambda}. \tag{3.8}$$

The point $P = \sigma^{-1}(u)$ is characterized by these equations and the fact that P is on the sphere. Thus, substituting (3.8) into $|\xi|^2 + \tau^2 = R^2$ gives

$$\frac{|u|^2}{\lambda^2} + R^2 \frac{(\lambda - 1)^2}{\lambda^2} = R^2,$$

from which we conclude

$$\lambda = \frac{|u|^2 + R^2}{2R^2}.$$

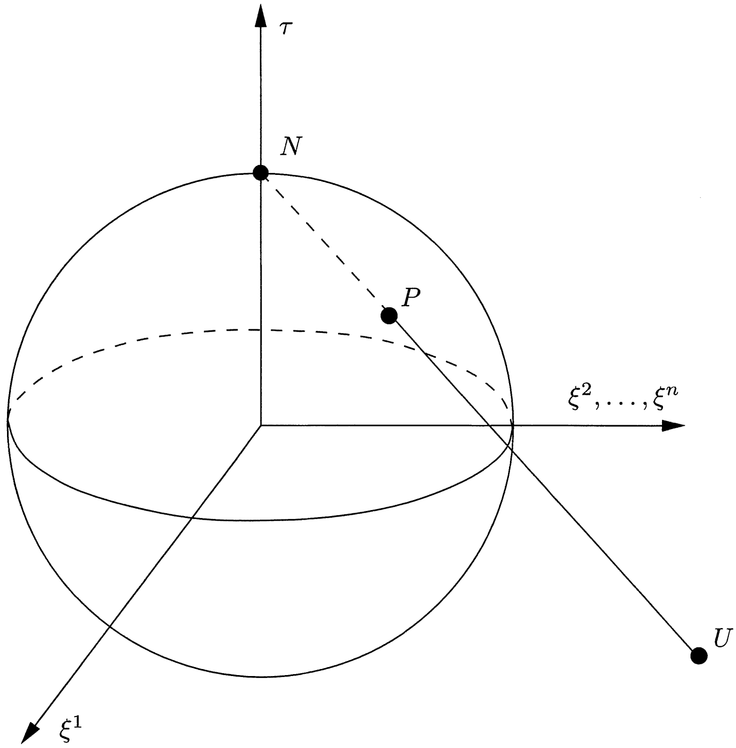


FIGURE 3.3. Stereographic projection.

Inserting this back into (3.8) gives the formula

$$\sigma^{-1}(u) = (\xi, \tau) = \left(\frac{2R^2 u}{|u|^2 + R^2}, R\frac{|u|^2 - R^2}{|u|^2 + R^2} \right), \tag{3.9}$$

which by construction maps $\mathbf{R}^n$ back to $\mathbf{S}^n_R - \{N\}$ and shows that σ is a diffeomorphism.

Lemma 3.4. *Stereographic projection is a conformal equivalence between* $\mathbf{S}^n_R - \{N\}$ *and* $\mathbf{R}^n$.

Proof. The inverse map σ^{-1} is a local parametrization, so we will use it to compute the pullback metric. Consider an arbitrary point $q \in \mathbf{R}^n$ and a vector $V \in T_q\mathbf{R}^n$, and compute

$$(\sigma^{-1})^* \mathring{g}_R(V, V) = \mathring{g}_R(\sigma_*^{-1}V, \sigma_*^{-1}V) = \bar{g}(\sigma_*^{-1}V, \sigma_*^{-1}V),$$

where $\bar{g}$ denotes the Euclidean metric on $\mathbf{R}^{n+1}$. Writing $V = V^i\partial_i$ and $\sigma^{-1}(u) = (\xi(u), \tau(u))$, the usual formula for the push-forward of a vector

can be written

$$\begin{aligned}
\sigma_*^{-1}V &= V^i \frac{\partial \xi^j}{\partial u^i}\frac{\partial}{\partial \xi^j} + V^i \frac{\partial \tau}{\partial u^i}\frac{\partial}{\partial \tau} \\
&= V\xi^j \frac{\partial}{\partial \xi^j} + V\tau \frac{\partial}{\partial \tau}.
\end{aligned}$$

Now

$$\begin{aligned}
V\xi^j &= V\left(\frac{2R^2u^j}{|u|^2+R^2}\right) \\
&= \frac{2R^2V^j}{|u|^2+R^2} - \frac{4R^2u^j\langle V,u\rangle}{(|u|^2+R^2)^2}; \\
V\tau &= V\left(R\frac{|u|^2-R^2}{|u|^2+R^2}\right) \\
&= \frac{2R\langle V,u\rangle}{|u|^2+R^2} - \frac{2R(|u|^2-R^2)\langle V,u\rangle}{(|u|^2+R^2)^2} \\
&= \frac{4R^3\langle V,u\rangle}{(|u|^2+R^2)^2},
\end{aligned}$$

where we have used the notation $V(|u|^2) = 2\sum_k V^k u^k = 2\langle V,u\rangle$. Therefore,

$$\begin{aligned}
\bar{g}(\sigma_*^{-1}V, \sigma_*^{-1}V) &= \sum_{j=1}^{n}(V\xi^j)^2 + (V\tau)^2 \\
&= \frac{4R^4|V|^2}{(|u|^2+R^2)^2} - \frac{16R^4\langle V,u\rangle^2}{(|u|^2+R^2)^3} + \frac{16R^4|u|^2\langle V,u\rangle^2}{(|u|^2+R^2)^4} \\
&\quad + \frac{16R^6\langle V,u\rangle^2}{(|u|^2+R^2)^4} \\
&= \frac{4R^4|V|^2}{(|u|^2+R^2)^2}.
\end{aligned}$$

In other words,

$$(\sigma^{-1})^*\mathring{g}_R = \frac{4R^4}{(|u|^2+R^2)^2}\bar{g}, \tag{3.10}$$

where now $\bar{g}$ represents the Euclidean metric on $\mathbf{R}^n$, and so σ is a conformal equivalence. □

It follows immediately from this lemma that the sphere is *locally conformally flat;* i.e., each point $p \in \mathbf{S}^n_R$ has a neighborhood that is conformally equivalent to an open set in $\mathbf{R}^n$. Stereographic projection gives such an equivalence for a neighborhood of any point except the north pole; applying a suitable rotation and then stereographic projection (or stereographic projection from the south pole), we get such an equivalence for a neighborhood of the north pole as well.

Hyperbolic Spaces

Our third class of model Riemannian manifolds is the hyperbolic spaces of dimension n. For each $R > 0$ we will describe a homogeneous, isotropic Riemannian manifold $\mathbf{H}^n_R$, called *hyperbolic space of radius R,* analogous to the sphere of radius R. The special case $R = 1$ is denoted $\mathbf{H}^n$ and is called simply *hyperbolic space.* There are three equivalent models of the hyperbolic spaces, each of which is useful in certain contexts. We'll introduce all of them and show that they are isometric.

Proposition 3.5. *For any fixed $R > 0$, the following Riemannian manifolds are all mutually isometric.*

(*a*) (Hyperboloid model) $\mathbf{H}^n_R$ *is the "upper sheet" $\{\tau > 0\}$ of the two-sheeted hyperboloid in $\mathbf{R}^{n+1}$ defined in coordinates $(\xi^1, \dots, \xi^n, \tau)$ by the equation $\tau^2 - |\xi|^2 = R^2$, with the metric*

$$h^1_R = \iota^* m,$$

where $\iota\colon \mathbf{H}^n_R \to \mathbf{R}^{n+1}$ is inclusion, and m is the Minkowski metric (3.5) *on $\mathbf{R}^{n+1}$.*

(*b*) (Poincaré ball model) $\mathbf{B}^n_R$ *is the ball of radius R in $\mathbf{R}^n$, with the metric given in coordinates $(u^1, \dots, u^n)$ by*

$$h^2_R = 4R^4 \frac{(du^1)^2 + \cdots + (du^n)^2}{(R^2 - |u|^2)^2}.$$

(*c*) (Poincaré half-space model) $\mathbf{U}^n_R$ *is the upper half-space in $\mathbf{R}^n$ defined in coordinates $(x^1, \dots, x^{n-1}, y)$ by $\{y > 0\}$, with the metric*

$$h^3_R = R^2 \frac{(dx^1)^2 + \cdots + (dx^{n-1})^2 + dy^2}{y^2}.$$

Proof. We begin by giving a geometric construction of a diffeomorphism

$$\pi\colon \mathbf{H}^n_R \to \mathbf{B}^n_R$$

from the hyperboloid to the ball, which we call *hyperbolic stereographic projection,* and which turns out to be an isometry between the two metrics given in (a) and (b).

Let $S \in \mathbf{R}^{n+1}$ denote the point $S = (0, \dots, 0, -R)$. For any $P = (\xi^1, \dots, \xi^n, \tau) \in \mathbf{H}^n_R \subset \mathbf{R}^{n+1}$, set $\pi(P) = u \in \mathbf{B}^n_R$, where $U = (u, 0) \in \mathbf{R}^{n+1}$ is the point where the line through S and P intersects the hyperplane $\{\tau = 0\}$ (Figure 3.4). U is characterized by $\overrightarrow{SU} = \lambda \overrightarrow{SP}$ for some nonzero scalar λ, or

$$\begin{aligned} u^i &= \lambda \xi^i, \\ R &= \lambda(\tau + R). \end{aligned} \tag{3.11}$$

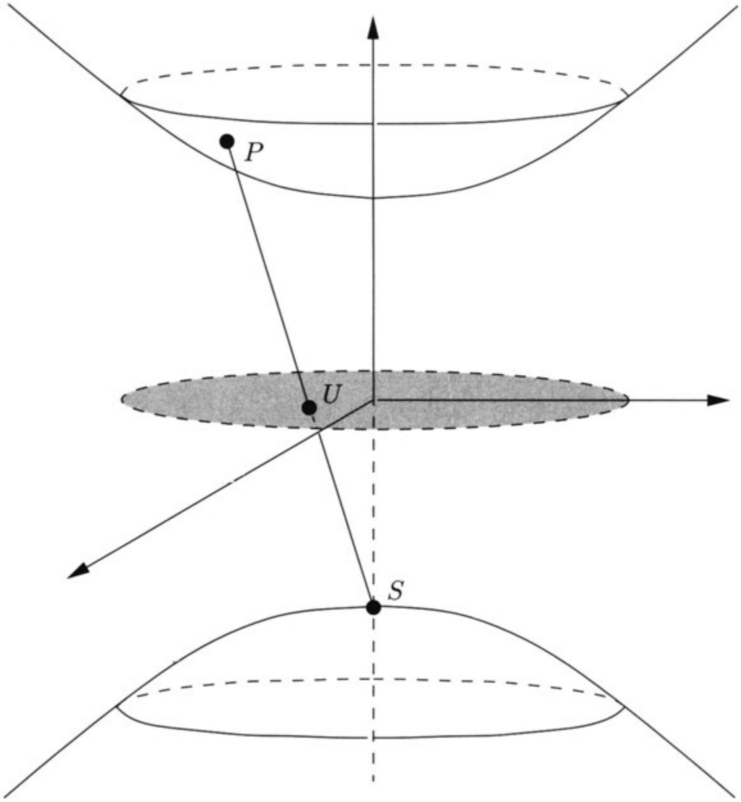


FIGURE 3.4. Hyperbolic stereographic projection.

These equations can be solved in the same manner as in the spherical case to yield

$$\pi(\xi,\tau) = u = \frac{R\xi}{R+\tau},$$

and its inverse map

$$\pi^{-1}(u) = (\xi,\tau) = \left(\frac{2R^2u}{R^2-|u|^2}, R\frac{R^2+|u|^2}{R^2-|u|^2}\right).$$

We will show that $(\pi^{-1})^* h_R^1 = h_R^2$. As before, let $V \in T_q\mathbf{B}_R^n$ and compute

$$(\pi^{-1})^* h_R^1(V,V) = h_R^1(\pi_*^{-1}V, \pi_*^{-1}V) = m(\pi_*^{-1}V, \pi_*^{-1}V).$$

The computation proceeds just as before. In this case, the relevant equations are

$$\begin{aligned}
V\xi^j &= \frac{2R^2V^j}{R^2-|u|^2} + \frac{4R^2u^j\langle V,u\rangle}{(R^2-|u|^2)^2};\\
V\tau &= \frac{4R^3\langle V,u\rangle}{(R^2-|u|^2)^2};\\
m(\pi_*^{-1}V,\pi_*^{-1}V) &= \sum_{j=1}^{n}(V\xi^j)^2 - (V\tau)^2\\
&= \frac{4R^4|V|^2}{(R^2-|u|^2)^2}\\
&= h_R^2(V,V).
\end{aligned}$$

Incidentally, this argument also shows that h_R^1 is positive definite, and thus is indeed a Riemannian metric, a fact that was not evident from the defining formula due to the fact that m is not positive definite.

Next we consider the Poincaré half-space model, by constructing an explicit diffeomorphism

$$\kappa\colon \mathbf{B}_R^n \to \mathbf{U}_R^n.$$

In this case it is more convenient to write the coordinates on the ball as $(u^1,\ldots,u^{n-1},v) = (u,v)$. In the 2-dimensional case, κ is easy to write down in complex notation $w=u+iv$ and $z=x+iy$. It is a variant of the classical Cayley transform:

$$\kappa(w) = z = -iR\frac{w+iR}{w-iR}. \tag{3.12}$$

It is shown in elementary complex analysis courses that this is a complex-analytic diffeomorphism taking $\mathbf{B}_R^2$ onto $\mathbf{U}_R^2$. Separating z into real and imaginary parts, this can also be written in real terms as

$$\kappa(u,v) = (x,y) = \left(\frac{2R^2u}{|u|^2+(v-R)^2}, R\frac{R^2-|u|^2-v^2}{|u|^2+(v-R)^2}\right).$$

This same formula makes sense in any dimension, and obviously maps the ball $\{|u|^2+v^2<R^2\}$ into the upper half-space. It is straightforward to check that its inverse is

$$\kappa^{-1}(x,y) = (u,v) = \left(\frac{2R^2x}{|x|^2+(y+R)^2}, R\frac{|x|^2+|y|^2-R^2}{|x|^2+(y+R)^2}\right),$$

so κ is a diffeomorphism, called the *generalized Cayley transform.* The verification that $\kappa^*h_R^3 = h_R^2$ is basically a long calculation, and is left to the reader. □

Exercise 3.11. Prove that $\kappa^* h^3_R = h^2_R$. Here are three different ways you might wish to proceed:

(i) Compute $h^3_R(\kappa_* V, \kappa_* V)$ directly, as in the proof of Proposition 3.5.

(ii) Show that κ is the restriction to the ball of the map $\sigma \circ \rho \circ \sigma^{-1}$, where $\sigma\colon \mathbf{S}^n_R \to \mathbf{R}^n$ is stereographic projection and $\rho\colon \mathbf{S}^n_R \to \mathbf{S}^n_R$ is the 90° rotation

$$\rho(\xi^1, \ldots, \xi^{n-1}, \xi^n, \tau) = (\xi^1, \ldots, \xi^{n-1}, -\tau, \xi^n),$$

taking the hemisphere $\{\tau < 0\}$ to the hemisphere $\{\xi^n > 0\}$. This shows that κ is a conformal map, and therefore it suffices to show that $h^3_R(\kappa_* V, \kappa_* V) = h^2_R(V, V)$ for a single strategically chosen vector V at each point. Do this for $V = \partial/\partial v$.

(iii) If you know some complex analysis, first do the 2-dimensional case using the complex form (3.12) of κ: Compute the pullback in complex notation, by noting that

$$h^3_R = R^2 \frac{dz\, d\bar z}{(\operatorname{Im} z)^2}, \qquad h^2_R = 4R^4 \frac{dw\, d\bar w}{(R^2 - |w|^2)^2},$$

and using the fact that a holomorphic diffeomorphism $z = F(w)$ is a conformal map with $F^*(dz\, d\bar z) = |F'(w)|^2 dw\, d\bar w$. Then show that the computation of $h^3_R(\kappa_* V, \kappa_* V)$ in higher dimensions can be reduced to the 2-dimensional case, by conjugating κ with a suitable orthogonal transformation in $n-1$ variables.

We often use the generic notation $\mathbf{H}^n_R$ to refer to any one of the manifolds of Proposition 3.5, and h_R to refer to the corresponding metric, using whichever model is most convenient for the application we have in mind. For example, the form of the metric in either the ball model or the half-space model makes it clear that the hyperbolic metric is locally conformally flat; indeed, in either model, the identity map gives a global conformal equivalence with an open subset of Euclidean space.

The symmetries of $\mathbf{H}^n_R$ are most easily seen in the hyperboloid model. Let $O(n,1)$ denote the group of linear maps from $\mathbf{R}^{n+1}$ to itself that preserve the Minkowski metric. (This is called the *Lorentz group* in the physics literature.) Note that each element of $O(n,1)$ preserves the set $\{\tau^2 - |\xi|^2 = R^2\}$, which has two components determined by $\{\tau > 0\}$ and $\{\tau < 0\}$. We let $O_+(n,1)$ denote the subgroup of $O(n,1)$ consisting of maps that take the component $\{\tau > 0\}$ to itself. Clearly $O_+(n,1)$ preserves $\mathbf{H}^n_R$, and because it preserves m it acts on $\mathbf{H}^n_R$ as isometries.

Proposition 3.6. *$O_+(n,1)$ acts transitively on the set of orthonormal bases on $\mathbf{H}^n_R$, and therefore $\mathbf{H}^n_R$ is homogeneous and isotropic.*

Proof. The argument is entirely analogous to the proof of Proposition 3.3, so we give only a sketch. If $p \in \mathbf{H}^n_R$ and $\{E_i\}$ is an orthonormal basis for $T_p\mathbf{H}^n_R$, an easy computation shows that $\{E_1, \ldots, E_n, E_{n+1} = p/R\}$ is a

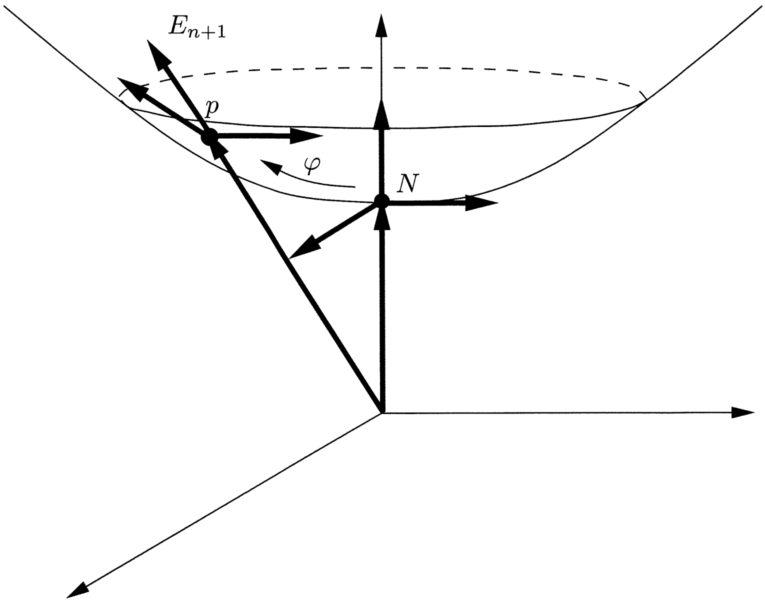


FIGURE 3.5. $O_+(n,1)$ acts transitively on orthonormal bases on $\mathbf{H}^n_R$.

basis for $\mathbf{R}^{n+1}$ such that m has the following expression in terms of the dual basis:

$$m = (\varphi^1)^2 + \cdots + (\varphi^n)^2 - (\varphi^{n+1})^2.$$

It follows easily that the matrix whose columns are the E_is is an element of $O_+(n,1)$ sending $N = (0,\ldots,0,R)$ to p and ∂_i to E_i (Figure 3.5). $\square$

Exercise 3.12. The spherical and hyperbolic metrics come in families $\mathring{g}_R$, h_R, parametrized by a positive real number R. We could have also defined a family of metrics on $\mathbf{R}^n$ by

$$\bar{g}_R = R^2 \delta_{ij} dx^i dx^j.$$

Why did we not bother?

Problems

3-1. Suppose $(\widetilde{M}, \tilde{g})$ is a Riemannian m-manifold, $M \subset \widetilde{M}$ is an embedded n-dimensional submanifold, and g is the induced Riemannian metric on M. For any point $p \in M$, show that there is a neighborhood $\widetilde{\mathcal{U}}$ of p in $\widetilde{M}$ and a smooth orthonormal frame $(E_1, \ldots, E_m)$ on $\widetilde{\mathcal{U}}$ such that $(E_1, \ldots, E_n)$ form an orthonormal basis for T_qM at each $q \in \widetilde{\mathcal{U}} \cap M$. Any such frame is called an *adapted orthonormal frame.* [Hint: Apply the Gram–Schmidt algorithm to the coordinate frame $\{\partial_i\}$ in slice coordinates.]

3-2. Suppose g is a pseudo-Riemannian metric on an n-manifold M. For any $p \in M$, show there is a smooth local frame $(E_1, \ldots, E_n)$ defined in a neighborhood of p such that g can be written locally in the form (3.4). Conclude that the index of g is constant on each component of M.

3-3. Let (M, g) be an oriented Riemannian manifold with volume element dV. The *divergence operator* $\operatorname{div}\colon \mathcal{T}(M) \to C^\infty(M)$ is defined by

$$d(i_X dV) = (\operatorname{div} X) dV,$$

where i_X denotes *interior multiplication* by X: for any k-form ω, $i_X\omega$ is the $(k-1)$-form defined by

$$i_X\omega(V_1, \ldots, V_{k-1}) = \omega(X, V_1, \ldots, V_{k-1}).$$

(a) Suppose M is a compact, oriented Riemannian manifold with boundary. Prove the following *divergence theorem* for $X \in \mathcal{T}(M)$:

$$\int_M \operatorname{div} X \, dV = \int_{\partial M} \langle X, N \rangle \, d\widetilde{V},$$

where N is the outward unit normal to ∂M and $d\widetilde{V}$ is the Riemannian volume element of the induced metric on ∂M.

(b) Show that the divergence operator satisfies the following product rule for a smooth function $u \in C^\infty(M)$:

$$\operatorname{div}(uX) = u \operatorname{div} X + \langle \operatorname{grad} u, X \rangle,$$

and deduce the following "integration by parts" formula:

$$\int_M \langle \operatorname{grad} u, X \rangle \, dV = -\int_M u \operatorname{div} X \, dV + \int_{\partial M} u \langle X, N \rangle \, d\widetilde{V}.$$

3-4. Let (M, g) be a compact, connected, oriented Riemannian manifold with boundary. For $u \in C^\infty(M)$, the *Laplacian* of u, denoted Δu, is defined to be the function $\Delta u = \operatorname{div}(\operatorname{grad} u)$. A function $u \in C^\infty(M)$ is said to be *harmonic* if $\Delta u = 0$.

(a) Prove *Green's identities:*

$$\int_M u\Delta v\, dV + \int_M \langle \operatorname{grad} u, \operatorname{grad} v\rangle\, dV = \int_{\partial M} u\, Nv\, d\widetilde{V}.$$

$$\int_M (u\Delta v - v\Delta u)\, dV = \int_{\partial M} (u\, Nv - v\, Nu) d\widetilde{V}.$$

(b) If $\partial M \neq \emptyset$, and u, v are harmonic functions on M whose restrictions to ∂M agree, show that $u \equiv v$.

(c) If $\partial M = \emptyset$, show that the only harmonic functions on M are the constants.

3-5. Let M be a compact oriented Riemannian manifold (without boundary). A real number λ is called an *eigenvalue* of the Laplacian if there exists a smooth function u on M, not identically zero, such that $\Delta u = \lambda u$. In this case, u is called an *eigenfunction* corresponding to λ.

(a) Prove that 0 is an eigenvalue of Δ, and that all other eigenvalues are strictly negative.

(b) If u and v are eigenfunctions corresponding to distinct eigenvalues, show that $\int_M uv\, dV = 0$.

3-6. Consider $\mathbf{R}^n$ as a Riemannian manifold with the Euclidean metric.

(a) Let $E(n)$ be the set of $(n+1) \times (n+1)$ real matrices of the form

$$\begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix},$$

where $A \in O(n)$ and $b \in \mathbf{R}^n$ (considered as a column vector). Show that $E(n)$ is a closed Lie subgroup of $GL(n+1, \mathbf{R})$, called the *Euclidean group* or the *group of rigid motions.*

(b) Define a map $E(n) \times \mathbf{R}^n \to \mathbf{R}^n$ by identifying $\mathbf{R}^n$ with the subset

$$S = \{(x, 1) \in \mathbf{R}^{n+1} : x \in \mathbf{R}^n\}$$

of $\mathbf{R}^{n+1}$ and restricting the linear action of $E(n)$ on $\mathbf{R}^{n+1}$ to S. Show that this is a smooth action of $E(n)$ on $\mathbf{R}^n$ by isometries of the Euclidean metric.

(c) Show that $E(n)$ acts transitively on $\mathbf{R}^n$, and takes any orthonormal basis to any other one, so Euclidean space is homogeneous and isotropic.

3-7. Let $\mathbf{U}^2$ denote the hyperbolic plane, i.e., the upper half-plane in $\mathbf{R}^2$ with the metric $h = (dx^2 + dy^2)/y^2$. Let $SL(2, \mathbf{R})$ denote the group of 2×2 real matrices of determinant 1.

(a) Considering $\mathbf{U}^2$ as a subset of the complex plane with coordinate $z = x + iy$, let

$$A \cdot z = \frac{az+b}{cz+d}, \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{R}).$$

Show that this defines a smooth action of $SL(2, \mathbf{R})$ on $\mathbf{U}^2$ by isometries of the hyperbolic metric.

(b) We have seen that $O_+(2,1)$ also acts on $\mathbf{U}^2$ by isometries. Show that $SL(2, \mathbf{R})/\{\pm I\} \cong SO_+(2,1)$, where $SO_+(2,1) = O_+(2,1) \cap SL(3, \mathbf{R})$.

3-8. Suppose $\widetilde{M}$ and M are smooth manifolds, and $p \colon \widetilde{M} \to M$ is a surjective submersion. For any $y \in M$, the *fiber* over y, denoted $\widetilde{M}_y$, is the inverse image $p^{-1}(y) \subset \widetilde{M}$; it is a closed, embedded submanifold by the implicit function theorem. If $\widetilde{M}$ has a Riemannian metric $\tilde{g}$, at each point $x \in \widetilde{M}$ the tangent space $T_x\widetilde{M}$ decomposes into an orthogonal direct sum

$$T_x\widetilde{M} = H_x \oplus V_x,$$

where $V_x := \operatorname{Ker} p_* = T_x\widetilde{M}_{p(x)}$ is the *vertical space* and $H_x := V_x^{\perp}$ is the *horizontal space.* If g is a Riemannian metric on M, p is said to be a *Riemannian submersion* if $\tilde{g}(X, Y) = g(p_*X, p_*Y)$ whenever X and Y are horizontal.

(a) Show that any vector field W on $\widetilde{M}$ can be written uniquely as $W = W^H + W^V$, where W^H is horizontal, W^V is vertical, and both W^H and W^V are smooth.

(b) If X is a vector field on M, show there is a unique smooth horizontal vector field $\widetilde{X}$ on $\widetilde{M}$, called the *horizontal lift* of X, that is p-related to X. (This means $p_*\widetilde{X}_q = X_{p(q)}$ for each $q \in \widetilde{M}$.)

(c) Let G be a Lie group acting smoothly on $\widetilde{M}$ by isometries of $\tilde{g}$, and suppose that $p \circ \varphi = p$ for all $\varphi \in G$ and that G acts transitively on each fiber $\widetilde{M}_y$. Show that there is a unique Riemannian metric g on M such that p is a Riemannian submersion. [Hint: First show that $\varphi_* V_x = V_{\varphi(x)}$ for any $\varphi \in G$.]

3-9. The *complex projective space* of dimension n, denoted $\mathbf{CP}^n$, is defined as the set of 1-dimensional complex subspaces of $\mathbf{C}^{n+1}$. Let $\pi\colon \mathbf{C}^{n+1} - \{0\} \to \mathbf{CP}^n$ denote the quotient map.

(a) Show that $\mathbf{CP}^n$ can be uniquely given the structure of a smooth, compact, real $2n$-dimensional manifold on which the Lie group $U(n+1)$ acts smoothly and transitively.

(b) Show that the restriction of π to $\mathbf{S}^{2n+1} \subset \mathbf{C}^{n+1}$ is a surjective submersion.

(c) Using Problem 3-8, show that the round metric on $\mathbf{S}^{2n+1}$ descends to a homogeneous and isotropic Riemannian metric on $\mathbf{CP}^n$, called the *Fubini–Study metric.*

3-10. Let G be a Lie group with Lie algebra $\mathfrak{g}$. A Riemannian metric g on G is said to be *left-invariant* if it is invariant under all left translations: $L_p^* g = g$ for all $p \in G$. Similarly, g is *right-invariant* if it is invariant under all right translations, and *bi-invariant* if it is both left- and right-invariant.

(a) Show that a metric g is left-invariant if and only if the coefficients $g_{ij} := g(X_i, X_j)$ of g with respect to any left-invariant frame $\{X_i\}$ are constants.

(b) Show that the restriction map $g \mapsto g|_{T_eG}$ gives a bijection between left-invariant metrics on G and inner products on $\mathfrak{g}$.

3-11. Suppose G is a compact, connected Lie group with a left-invariant metric g, and let dV denote the Riemannian volume element of g. Show that dV is bi-invariant. [Hint: Show that $R_p^* dV$ is left-invariant and positively oriented, and is therefore equal to $\varphi(p)dV$ for some positive number $\varphi(p)$. Show that $\varphi\colon G \to \mathbf{R}^+$ is a Lie group homomorphism, so its image is a compact subgroup of $\mathbf{R}^+$.]

3-12. If G is a Lie group and $p \in G$, conjugation by p gives a Lie group automorphism $C_p\colon G \to G$, called an *inner automorphism,* by $C_p(q) = pqp^{-1}$. Let $\mathrm{Ad}_p := (C_p)_*\colon \mathfrak{g} \to \mathfrak{g}$ be the induced Lie algebra automorphism. It is easy to check that $C_{p_1} \circ C_{p_2} = C_{p_1 p_2}$, so $\mathrm{Ad}\colon G \times \mathfrak{g} \to \mathfrak{g}$ is a representation of G, called the *adjoint representation.*

(a) Show that an inner product on $\mathfrak{g}$ induces a bi-invariant metric on G as in Problem 3-10 if and only if it is invariant under the adjoint representation.

(b) Show that every compact, connected Lie group admits a bi-invariant Riemannian metric. [Hint: Start with an arbitrary inner product $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}$ and integrate the function f defined by $f(p) := \langle \mathrm{Ad}_p X, \mathrm{Ad}_p Y \rangle$ over the group. You will need to use the result of Problem 3-11.]

4
Connections

Before we can define curvature on Riemannian manifolds, we need to study geodesics, the Riemannian generalizations of straight lines. It is tempting to define geodesics as curves that minimize length, at least between nearby points. However, this property turns out to be technically difficult to work with as a definition, so instead we'll choose a different property of straight lines and generalize that.

A curve in Euclidean space is a straight line if and only if its acceleration is identically zero. This is the property that we choose to take as a defining property of geodesics on a Riemannian manifold. To make sense of this idea, we're going to have to introduce a new object on manifolds, called a connection—essentially a coordinate-invariant set of rules for taking directional derivatives of vector fields.

We begin this chapter by examining more closely the problem of finding an invariant interpretation for the acceleration of a curve, as a way to motivate the definitions that follow. We then give a rather general definition of a connection, in terms of directional derivatives of sections of vector bundles. The special case in which the vector bundle is the tangent bundle is called a "linear connection," and it is on this case that we focus most of our attention. After deriving some basic properties of connections, we show how to use one to differentiate vector fields along curves, to define geodesics, and to "parallel translate" vector fields along curves.

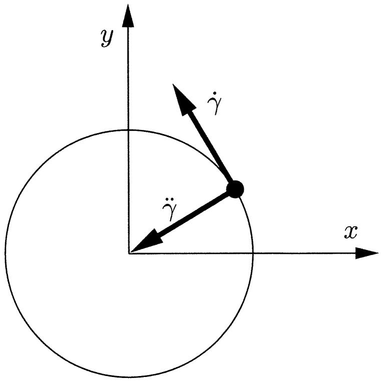


FIGURE 4.1. Euclidean coordinates.

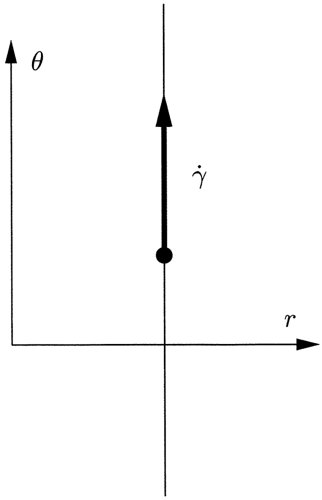


FIGURE 4.2. Polar coordinates.

The Problem of Differentiating Vector Fields

To see why we need a new kind of differentiation operator, consider a submanifold $M \subset \mathbf{R}^n$ with the induced Riemannian metric, and a smooth curve γ lying entirely in M. We want to think of a geodesic as a curve in M that is "as straight as possible." An intuitively plausible way to measure straightness is to compute the Euclidean acceleration $\ddot\gamma(t)$ as usual, and orthogonally project $\ddot\gamma(t)$ onto the tangent space $T_{\gamma(t)}M$. This yields a vector $\ddot\gamma(t)^\top$ tangent to M, the *tangential acceleration* of γ. We could then define a geodesic as a curve in M whose tangential acceleration is zero. This definition is easily seen to be invariant under rigid motions of $\mathbf{R}^n$, although at this point there is little reason to believe that it is an intrinsic invariant of M (one that depends only on the Riemannian geometry of M with its induced metric).

On an abstract Riemannian manifold, for which there is no "ambient Euclidean space" in which to differentiate, this technique is not available. Thus we have to find some way to make sense of the acceleration of a curve in an abstract manifold. Let $\gamma\colon (a,b) \to M$ be such a curve. As you know from your study of smooth manifold theory, the velocity vector $\dot\gamma(t)$ has a coordinate-independent meaning for each $t \in M$, and its expression in any coordinate system matches the usual notion of velocity of a curve in $\mathbf{R}^n$: $\dot\gamma(t) = (\dot\gamma^1(t), \ldots, \dot\gamma^n(t))$. However, unlike the velocity, the acceleration vector has no such coordinate-invariant interpretation. For example, consider the parametrized circle in the plane given in Euclidean coordinates by $(x(t), y(t)) = (\cos t, \sin t)$ (Figure 4.1). Its acceleration at time t is the unit

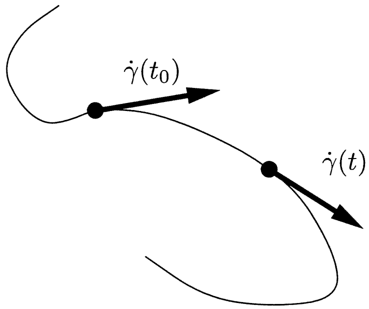


FIGURE 4.3. $\dot\gamma(t_0)$ and $\dot\gamma(t)$ lie in different vector spaces.

vector $(\ddot{x}(t), \ddot{y}(t)) = (-\cos t, -\sin t)$. But in polar coordinates, the same curve is described by $(r(t), \theta(t)) = (1, t)$ (Figure 4.2). In these coordinates, the acceleration vector is $(\ddot{r}(t), \ddot{\theta}(t)) = (0, 0)$!

The problem is this: If we wanted to make sense of $\ddot\gamma(t_0)$ by differentiating $\dot\gamma(t)$ with respect to t, we would have to write a difference quotient involving the vectors $\dot\gamma(t)$ and $\dot\gamma(t_0)$; but these live in different vector spaces ($T_{\gamma(t)}M$ and $T_{\gamma(t_0)}M$ respectively), so it doesn't make sense to subtract them (Figure 4.3).

The velocity vector $\dot\gamma(t)$ is an example of a "vector field along a curve," a concept for which we will give a rigorous definition presently. To interpret the acceleration of a curve in a manifold, what we need is some coordinate-invariant way to differentiate vector fields along curves. To do so, we need a way to compare values of the vector field at different points, or, intuitively, to "connect" nearby tangent spaces. This is where a connection comes in: it will be an additional piece of data on a manifold, a rule for computing directional derivatives of vector fields.

Connections

It turns out to be easiest to define a connection first as a way of differentiating sections of vector bundles. Later we will adapt the definition to the case of vector fields along curves.

Let $\pi\colon E \to M$ be a vector bundle over a manifold M, and let $\mathcal{E}(M)$ denote the space of smooth sections of E. A *connection* in E is a map

$$\nabla\colon \mathcal{T}(M) \times \mathcal{E}(M) \to \mathcal{E}(M),$$

written $(X, Y) \mapsto \nabla_X Y$, satisfying the following properties:

(a) $\nabla_X Y$ is linear over $C^\infty(M)$ in X:

$$\nabla_{fX_1+gX_2} Y = f\nabla_{X_1} Y + g\nabla_{X_2} Y \qquad \text{for } f, g \in C^\infty(M);$$

(b) $\nabla_X Y$ is linear over $\mathbf{R}$ in Y:

$$\nabla_X(aY_1 + bY_2) = a\nabla_X Y_1 + b\nabla_X Y_2 \qquad \text{for } a, b \in \mathbf{R};$$

(c) ∇ satisfies the following product rule:

$$\nabla_X(fY) = f\nabla_X Y + (Xf)Y \qquad \text{for } f \in C^\infty(M).$$

The symbol ∇ is read "del," and $\nabla_X Y$ is called the *covariant derivative of* Y *in the direction of* X.

Although a connection is defined by its action on global sections, it follows from the definitions that it is actually a local operator, as the next lemma shows.

Lemma 4.1. *If* ∇ *is a connection in a bundle* E, $X \in \mathcal{T}(M)$, $Y \in \mathcal{E}(M)$, *and* $p \in M$, *then* $\nabla_X Y|_p$ *depends only on the values of* X *and* Y *in an arbitrarily small neighborhood of* p. *More precisely, if* $X = \widetilde{X}$ *and* $Y = \widetilde{Y}$ *on a neighborhood of* p, *then* $\nabla_X Y|_p = \nabla_{\widetilde{X}} \widetilde{Y}|_p$.

Proof. First consider Y. Replacing Y by $Y - \widetilde{Y}$, it clearly suffices to show that $\nabla_X Y|_p = 0$ if Y vanishes on a neighborhood U of p.

Choose a bump function $\varphi \in C^\infty(M)$ with support in U such that $\varphi(p) = 1$. The hypothesis that Y vanishes on U implies that $\varphi Y \equiv 0$ on all of M, so $\nabla_X(\varphi Y) = \nabla_X(0 \cdot \varphi Y) = 0\nabla_X(\varphi Y) = 0$. Thus for any $X \in \mathcal{T}(M)$, the product rule gives

$$0 = \nabla_X(\varphi Y) = (X\varphi)Y + \varphi(\nabla_X Y). \tag{4.1}$$

Now $Y \equiv 0$ on the support of φ, so the first term on the right is identically zero. Evaluating (4.1) at p shows that $\nabla_X Y|_p = 0$. The argument for X is similar but easier. □

Exercise 4.1. Complete the proof of Lemma 4.1 by showing that $\nabla_X Y$ and $\nabla_{\widetilde{X}} Y$ agree at p if $X = \widetilde{X}$ on a neighborhood of p.

The preceding lemma tells us that we can compute $\nabla_X Y$ at p knowing only the values of X and Y near p. In fact, as the next lemma shows, we need only know the value of X at p itself.

Lemma 4.2. *With notation as in Lemma 4.1,* $\nabla_X Y|_p$ *depends only on the values of* Y *in a neighborhood of* p *and the value of* X *at* p.

Proof. By linearity, it suffices to show that $\nabla_X Y|_p = 0$ whenever $X_p = 0$. Choose a coordinate neighborhood U of p, and write $X = X^i \partial_i$ in coordinates on U, with $X^i(p) = 0$. Then, for any $Y \in \mathcal{E}(M)$,

$$\nabla_X Y|_p = \nabla_{X^i \partial_i} Y|_p = X^i(p) \nabla_{\partial_i} Y|_p = 0.$$

In the first equality, we used Lemma 4.1, which allows us to evaluate $\nabla_X Y|_p$ by computing locally in U; in the second, we used linearity of $\nabla_X Y$ over $C^\infty(M)$ in X. □

Because of Lemma 4.2, we can write $\nabla_{X_p}Y$ in place of $\nabla_X Y|_p$. This can be thought of as a directional derivative of Y at p in the direction of the vector X_p.

Linear Connections

Now we specialize to connections in the tangent bundle of a manifold. A *linear connection* on M is a connection in TM, i.e., a map

$$\nabla\colon \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$$

satisfying properties (a)–(c) in the definition of a connection above.

A linear connection on M is often simply called a *connection on M*. (The term *affine connection* is also frequently used synonymously with linear connection, although some authors make a subtle distinction between the two terms; cf., for example, [KN63, volume 1].)

Although the definition of a linear connection resembles the characterization of $\binom{2}{1}$-tensor fields given by the tensor characterization lemma (Lemma 2.4), a linear connection is not a tensor field because it is not linear over $C^\infty(M)$ in Y, but instead satisfies the product rule.

Next we examine how a linear connection appears in components. Let $\{E_i\}$ be a local frame for TM on an open subset $U \subset M$. We will usually work with a coordinate frame $E_i = \partial_i$, but it is useful to start by doing the computations for more general frames. For any choices of the indices i and j, we can expand $\nabla_{E_i}E_j$ in terms of this same frame:

$$\nabla_{E_i}E_j = \Gamma_{ij}^k E_k. \tag{4.2}$$

This defines n^3 functions Γ_{ij}^k on U, called the *Christoffel symbols* of ∇ with respect to this frame. The following lemma shows that the action of the connection ∇ on U is completely determined by its Christoffel symbols.

Lemma 4.3. *Let ∇ be a linear connection, and let $X, Y \in \mathcal{T}(U)$ be expressed in terms of a local frame by $X = X^iE_i$, $Y = Y^jE_j$. Then*

$$\nabla_X Y = (XY^k + X^iY^j\Gamma_{ij}^k)E_k. \tag{4.3}$$

Proof. Just use the defining rules for a connection and compute:

$$\begin{aligned}\nabla_X Y &= \nabla_X(Y^jE_j)\\ &= (XY^j)E_j + Y^j\nabla_{X^iE_i}E_j\\ &= (XY^j)E_j + X^iY^j\nabla_{E_i}E_j\\ &= XY^jE_j + X^iY^j\Gamma_{ij}^kE_k.\end{aligned}$$

Renaming the dummy index in the first term yields (4.3). □

Existence of Connections

So far, we have studied properties of connections, but have not produced any, so you might be wondering if they are plentiful or rare. In fact, they are quite plentiful, as we will show shortly. Let's begin with a trivial example: on $\mathbf{R}^n$, define the *Euclidean connection* by

$$\overline{\nabla}_X \left(Y^j \partial_j\right) = (XY^j)\partial_j. \tag{4.4}$$

In other words, $\overline{\nabla}_X Y$ is just the vector field whose components are the ordinary directional derivatives of the components of Y in the direction X. It is easy to check that this satisfies the required properties for a connection, and that its Christoffel symbols in standard coordinates are all zero. In fact, there are many more connections on $\mathbf{R}^n$, or indeed on any manifold covered by a single coordinate chart; the following lemma shows how to construct all of them explicitly.

Lemma 4.4. *Suppose M is a manifold covered by a single coordinate chart. There is a one-to-one correspondence between linear connections on M and choices of n^3 smooth functions $\{\Gamma_{ij}^k\}$ on M, by the rule*

$$\nabla_X Y = \left(X^i \partial_i Y^k + X^i Y^j \Gamma_{ij}^k\right) \partial_k. \tag{4.5}$$

Proof. Observe that (4.5) is equivalent to (4.3) when $E_i = \partial_i$ is a coordinate frame, so for every connection the functions $\{\Gamma_{ij}^k\}$ defined by (4.2) satisfy (4.5). On the other hand, given $\{\Gamma_{ij}^k\}$, it is easy to see by inspection that (4.5) is smooth if X and Y are, linear over $\mathbf{R}$ in Y, and linear over $C^\infty(M)$ in X, so only the product rule requires checking; this is a straightforward computation left to the reader. □

Exercise 4.2. Complete the proof of Lemma 4.4.

Proposition 4.5. *Every manifold admits a linear connection.*

Proof. Cover M with coordinate charts $\{U_\alpha\}$; the preceding lemma guarantees the existence of a connection ∇^α on each U_α. Choosing a partition of unity $\{\varphi_\alpha\}$ subordinate to $\{U_\alpha\}$, we'd like to patch the ∇^αs together by the formula

$$\nabla_X Y = \sum_\alpha \varphi_\alpha \nabla_X^\alpha Y. \tag{4.6}$$

Again, it is obvious by inspection that this expression is smooth, linear over $\mathbf{R}$ in Y, and linear over $C^\infty(M)$ in X. We have to be a bit careful with the product rule, though, since a linear combination of connections is not necessarily a connection. (You can check, for example, that if ∇^1 and ∇^2

are connections, neither $\frac{1}{2}\nabla^1$ nor $\nabla^1 + \nabla^2$ satisfies the product rule.) By direct computation,

$$\begin{aligned}\nabla_X(fY) &= \sum_\alpha \varphi_\alpha \nabla^\alpha_X(fY)\\ &= \sum_\alpha \varphi_\alpha((Xf)Y + f\nabla^\alpha_X Y)\\ &= (Xf)Y + f\sum_\alpha \varphi_\alpha \nabla^\alpha_X Y\\ &= (Xf)Y + f\nabla_X Y.\end{aligned}$$

□

Covariant Derivatives of Tensor Fields

By definition, a linear connection on M is a way to compute covariant derivatives of vector fields. In fact, any linear connection automatically induces connections on all tensor bundles over M, and thus gives us a way to compute covariant derivatives of any tensor field.

Lemma 4.6. *Let ∇ be a linear connection on M. There is a unique connection in each tensor bundle $T^k_l M$, also denoted ∇, such that the following conditions are satisfied.*

(a) On TM, ∇ agrees with the given connection.

(b) On $T^0 M$, ∇ is given by ordinary differentiation of functions:

$$\nabla_X f = Xf.$$

(c) ∇ obeys the following product rule with respect to tensor products:

$$\nabla_X(F \otimes G) = (\nabla_X F) \otimes G + F \otimes (\nabla_X G).$$

(d) ∇ commutes with all contractions: if "tr" denotes the trace on any pair of indices,

$$\nabla_X(\operatorname{tr} Y) = \operatorname{tr}(\nabla_X Y).$$

This connection satisfies the following additional properties:

(i) ∇ obeys the following product rule with respect to the natural pairing between a covector field ω and a vector field Y:

$$\nabla_X \langle\omega, Y\rangle = \langle\nabla_X \omega, Y\rangle + \langle\omega, \nabla_X Y\rangle.$$

(*ii*) *For any* $F \in \mathcal{T}^k_l(M)$*, vector fields* Y_i*, and* 1*-forms* ω^j*,*

$$\begin{aligned}(\nabla_X F)(\omega^1, \ldots, \omega^l, Y_1, \ldots, Y_k) &= X(F(\omega^1, \ldots, \omega^l, Y_1, \ldots, Y_k)) \\ &\quad - \sum_{j=1}^{l} F(\omega^1, \ldots, \nabla_X \omega^j, \ldots, \omega^l, Y_1, \ldots, Y_k) \\ &\quad - \sum_{i=1}^{k} F(\omega^1, \ldots, \omega^l, Y_1, \ldots, \nabla_X Y_i, \ldots, Y_k).\end{aligned} \tag{4.7}$$

Exercise 4.3. Prove Lemma 4.6. [Hint: Show that the defining properties imply (i) and (ii); then use these to prove existence.]

Exercise 4.4. Let ∇ be a linear connection. If ω is a 1-form and X a vector field, show that the coordinate expression for $\nabla_X \omega$ is

$$\nabla_X \omega = \left(X^i \partial_i \omega_k - X^i \omega_j \Gamma^j_{ik} \right) dx^k,$$

where $\{\Gamma^k_{ij}\}$ are the Christoffel symbols of the given connection ∇ on TM. Find a coordinate formula for $\nabla_X F$, where $F \in \mathcal{T}^k_l(M)$ is a tensor field of any rank.

Because the covariant derivative $\nabla_X Y$ of a vector field (or tensor field) Y is linear over $C^\infty(M)$ in X, it can be used to construct another tensor field called the total covariant derivative, as follows.

Lemma 4.7. *If* ∇ *is a linear connection on* M*, and* $F \in \mathcal{T}^k_l(M)$*, the map* $\nabla F \colon \mathcal{T}^1(M) \times \cdots \times \mathcal{T}^1(M) \times \mathcal{T}(M) \times \cdots \times \mathcal{T}(M) \to C^\infty(M)$*, given by*

$$\nabla F(\omega^1, \ldots, \omega^l, Y_1, \ldots, Y_k, X) = \nabla_X F(\omega^1, \ldots, \omega^l, Y_1, \ldots, Y_k),$$

defines a $\binom{k+1}{l}$*-tensor field.*

Proof. This follows immediately from the tensor characterization lemma: $\nabla_X F$ is a tensor field, so it is multilinear over $C^\infty(M)$ in its $k+l$ arguments; and it is linear over $C^\infty(M)$ in X by definition of a connection. □

The tensor field ∇F is called the *total covariant derivative* of F. For example, let u be a smooth function on M. Then $\nabla u \in \mathcal{T}^1(M)$ is just the 1-form du, because both tensors have the same action on vectors: $\langle \nabla u, X\rangle = \nabla_X u = Xu = \langle du, X\rangle$. The 2-tensor $\nabla^2 u = \nabla(\nabla u)$ is called the *covariant Hessian of* u.

Exercise 4.5. Show that for any $u \in C^\infty(M)$ and $X, Y \in \mathcal{T}(M)$,

$$\nabla^2 u(X, Y) = Y(Xu) - (\nabla_Y X)u. \tag{4.8}$$

When we write the components of a total covariant derivative in terms of coordinates, we use a semicolon to separate indices resulting from differentiation from the preceding indices. Thus, for example, if Y is a vector field written in components as $Y = Y^i\partial_i$, the components of the $\binom{1}{1}$-tensor field ∇Y are written $Y^i{}_{;j}$, so that

$$\nabla Y = Y^i{}_{;j}\partial_i \otimes dx^j,$$

with

$$Y^i{}_{;j} = \partial_j Y^i + Y^k\Gamma^i_{jk}.$$

More generally, the next lemma gives a formula for the components of covariant derivatives of arbitrary tensor fields.

Lemma 4.8. *Let ∇ be a linear connection. The components of the total covariant derivative of a $\binom{k}{l}$-tensor field F with respect to a coordinate system are given by*

$$F^{j_1\ldots j_l}_{i_1\ldots i_k;m} = \partial_m F^{j_1\ldots j_l}_{i_1\ldots i_k} + \sum_{s=1}^{l} F^{j_1\ldots p\ldots j_l}_{i_1\ldots i_k}\Gamma^{j_s}_{mp} - \sum_{s=1}^{k} F^{j_1\ldots j_l}_{i_1\ldots p\ldots i_k}\Gamma^p_{mi_s}.$$

Exercise 4.6. Prove Lemma 4.8.

Vector Fields Along Curves

Without further qualification, a *curve* in a manifold M always means for us a smooth, parametrized curve; that is, a smooth map $\gamma\colon I \to M$, where $I \subset \mathbf{R}$ is some interval. Unless otherwise specified, we won't worry about whether the interval is open or closed, bounded or unbounded. A *curve segment* is a curve whose domain is a closed, bounded interval $[a, b] \subset \mathbf{R}$.

If $\gamma\colon I \to M$ is a curve and the interval I has an endpoint, smoothness of γ means by definition that γ extends to a smooth curve defined on some open interval containing I. It can be shown (though we will not do so) that this notion of smoothness is equivalent to the component functions γ^i in any local coordinates having one-sided derivatives of all orders at the endpoint, or having derivatives of all orders that extend continuously to the endpoint. When working with a smooth curve γ defined on an interval that has one or two endpoints, we can always extend γ to a smooth curve on a slightly larger open interval, work with that curve, and restrict back to the original interval; the values on I of any continuous function of the derivatives of γ are independent of the extension. Thus in proofs we can assume whenever convenient that γ is defined on an open interval.

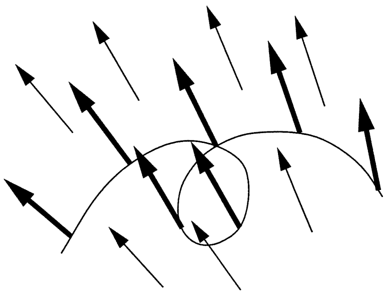


FIGURE 4.4. Extendible vector field.

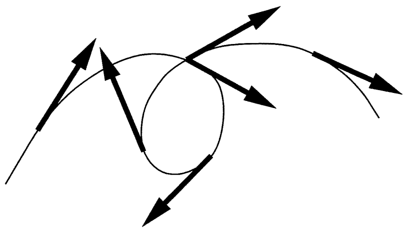


FIGURE 4.5. Nonextendible vector field.

Let $\gamma\colon I \to M$ be a curve. At any time $t \in I$, the *velocity* $\dot\gamma(t)$ of γ is invariantly defined as the push-forward $\gamma_*(d/dt)$. It acts on functions by

$$\dot\gamma(t) f = \frac{d}{dt}(f \circ \gamma)(t).$$

As mentioned above, this corresponds to the usual notion of velocity in coordinates. If we write the coordinate representation of γ as $\gamma(t) = (\gamma^1(t), \dots, \gamma^n(t))$, then

$$\dot\gamma(t) = \dot\gamma^i(t)\partial_i. \tag{4.9}$$

(A dot always denotes the ordinary derivative with respect to t.)

A *vector field along a curve* $\gamma\colon I \to M$ is a smooth map $V\colon I \to TM$ such that $V(t) \in T_{\gamma(t)}M$ for every $t \in I$. We let $\mathcal{T}(\gamma)$ denote the space of vector fields along γ. The most obvious example of a vector field along a curve γ is its velocity vector: $\dot\gamma(t) \in T_{\gamma(t)}M$ for each t, and the coordinate expression (4.9) shows that it is smooth. Here is another example: If γ is a curve in $\mathbf{R}^2$, let $N(t) = J\dot\gamma(t)$, where J is counterclockwise rotation by $\pi/2$, so $N(t)$ is normal to $\dot\gamma(t)$. In components, $N(t) = (-\dot\gamma^2(t), \dot\gamma^1(t))$, so N is a smooth vector field along γ.

A large class of examples is provided by the following construction: Suppose $\gamma\colon I \to M$ is a curve, and $\widetilde V \in \mathcal{T}(M)$ is a vector field on M. For each $t \in I$, let $V(t) = \widetilde V_{\gamma(t)}$. It is easy to check in coordinates that V is smooth. A vector field V along γ is said to be *extendible* if there exists a vector field $\widetilde V$ on a neighborhood of the image of γ that is related to V in this way (Figure 4.4). Not every vector field along a curve need be extendible; for example, if $\gamma(t_1) = \gamma(t_2)$ but $\dot\gamma(t_1) \neq \dot\gamma(t_2)$ (Figure 4.5), then $\dot\gamma$ is not extendible.

Covariant Derivatives Along Curves

Now we can address the question that originally motivated the definition of connections: How can we make sense of the directional derivative of a vector field along a curve?

Lemma 4.9. *Let ∇ be a linear connection on M. For each curve $\gamma\colon I \to M$, ∇ determines a unique operator*

$$D_t\colon \mathcal{T}(\gamma) \to \mathcal{T}(\gamma)$$

satisfying the following properties:

(*a*) *Linearity over* **R**:

$$D_t(aV + bW) = aD_tV + bD_tW \qquad \text{for } a, b \in \mathbf{R}.$$

(*b*) *Product rule:*

$$D_t(fV) = \dot{f}V + fD_tV \qquad \text{for } f \in C^\infty(I).$$

(*c*) *If V is extendible, then for any extension $\widetilde{V}$ of V,*

$$D_tV(t) = \nabla_{\dot\gamma(t)}\widetilde{V}.$$

For any $V \in \mathcal{T}(\gamma)$, D_tV is called the *covariant derivative of V along γ.*

Proof. First we show uniqueness. Suppose D_t is such an operator, and let $t_0 \in I$ be arbitrary. An argument similar to that of Lemma 4.1 shows that the value of D_tV at t_0 depends only on the values of V in any interval $(t_0 - \varepsilon, t_0 + \varepsilon)$ containing t_0. (If I has an endpoint, extend γ to a slightly bigger open interval, prove the lemma there, and then restrict back to I.)

Choose coordinates near $\gamma(t_0)$, and write

$$V(t) = V^j(t)\partial_j$$

near t_0. Then by the properties of D_t, since ∂_j is extendible,

$$\begin{aligned} D_tV(t_0) &= \dot V^j(t_0)\partial_j + V^j(t_0)\nabla_{\dot\gamma(t_0)}\partial_j \\ &= \left(\dot V^k(t_0) + V^j(t_0)\dot\gamma^i(t_0)\Gamma^k_{ij}(\gamma(t_0))\right)\partial_k. \end{aligned} \tag{4.10}$$

This shows that such an operator is unique if it exists.

For existence, if $\gamma(I)$ is contained in a single chart, we can *define* D_tV by (4.10); the easy verification that it satisfies the requisite properties is left to the reader. In the general case, we can cover $\gamma(I)$ with coordinate charts and define D_tV by this formula in each chart, and uniqueness implies the various definitions agree whenever two or more charts overlap. □

Exercise 4.7. Improve Lemma 4.1 by showing that $\nabla_{X_p}Y$ actually depends only on the values of Y along any curve tangent to X_p. More precisely, suppose that $\gamma\colon (-\varepsilon,\varepsilon) \to M$ is a curve with $\gamma(0) = p$ and $\dot\gamma(0) = X_p$, and suppose Y and $\widetilde{Y}$ are vector fields that agree along γ. Show that $\nabla_{X_p}Y = \nabla_{X_p}\widetilde{Y}$.

Geodesics

Armed with the notion of covariant differentiation along curves, we can now define acceleration and geodesics.

Let M be a manifold with a linear connection ∇, and let γ be a curve in M. The *acceleration* of γ is the vector field $D_t\dot\gamma$ along γ. A curve γ is called a *geodesic* with respect to ∇ if its acceleration is zero: $D_t\dot\gamma \equiv 0$.

Exercise 4.8. Show that the geodesics on $\mathbf{R}^n$ with respect to the Euclidean connection (4.4) are exactly the straight lines with constant speed parametrizations.

Theorem 4.10. (Existence and Uniqueness of Geodesics) *Let M be a manifold with a linear connection. For any $p \in M$, any $V \in T_pM$, and any $t_0 \in \mathbf{R}$, there exist an open interval $I \subset \mathbf{R}$ containing t_0 and a geodesic $\gamma\colon I \to M$ satisfying $\gamma(t_0) = p$, $\dot\gamma(t_0) = V$. Any two such geodesics agree on their common domain.*

Proof. Choose coordinates (x^i) on some neighborhood U of p. From (4.10), a curve $\gamma\colon I \to U$ is a geodesic if and only if its component functions $\gamma(t) = (x^1(t),\dots,x^n(t))$ satisfy the *geodesic equation*

$$\ddot x^k(t) + \dot x^i(t)\dot x^j(t)\Gamma^k_{ij}(x(t)) = 0. \tag{4.11}$$

This is a second-order system of ordinary differential equations for the functions $x^i(t)$. The usual trick for proving existence and uniqueness for a second-order system is to introduce auxiliary variables $v^i = \dot x^i$ to convert it to the following equivalent first-order system in twice the number of variables:

$$\begin{aligned}\dot x^k(t) &= v^k(t),\\ \dot v^k(t) &= -v^i(t)v^j(t)\Gamma^k_{ij}(x(t)).\end{aligned}$$

By the existence and uniqueness theorem for first-order ODEs (see, for example, [Boo86, Theorem IV.4.1]), for any $(p, V) \in U \times \mathbf{R}^n$, there exist $\varepsilon > 0$ and a unique solution $\eta\colon (t_0 - \varepsilon, t_0 + \varepsilon) \to U \times \mathbf{R}^n$ to this system satisfying the initial condition $\eta(t_0) = (p, V)$. If we write the component functions of η as $\eta(t) = (x^i(t), v^i(t))$, then we can easily check that the

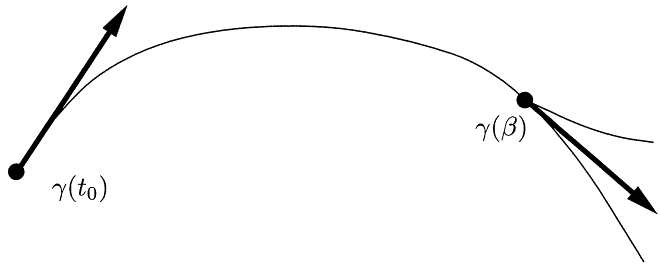


FIGURE 4.6. Uniqueness of geodesics.

curve $\gamma(t) = (x^1(t), \ldots, x^n(t))$ in U satisfies the existence claim of the lemma.

To prove the uniqueness claim, suppose $\gamma, \sigma \colon I \to M$ are geodesics defined on an open interval with $\gamma(t_0) = \sigma(t_0)$ and $\dot\gamma(t_0) = \dot\sigma(t_0)$. By the uniqueness part of the ODE theorem, they agree on some neighborhood of t_0. Let β be the supremum of numbers b such that they agree on $[t_0, b]$. If $\beta \in I$, then by continuity $\gamma(\beta) = \sigma(\beta)$ and $\dot\gamma(\beta) = \dot\sigma(\beta)$, and applying local uniqueness in a neighborhood of β, we conclude that they agree on a slightly larger interval (Figure 4.6), which is a contradiction. Arguing similarly to the left of t_0, we conclude that they agree on all of I. □

It follows from the uniqueness statement in the preceding theorem that for any $p \in M$ and $V \in T_pM$, there is a unique *maximal geodesic* (one that cannot be extended to any larger interval) $\gamma \colon I \to M$ with $\gamma(0) = p$ and $\dot\gamma(0) = V$, defined on some open interval I; just let I be the union of all open intervals on which such a geodesic is defined, and observe that the various geodesics agree where they overlap. This maximal geodesic is often called simply *the geodesic with initial point p and initial velocity V*, and is denoted γ_V. (The initial point p does not need to be specified in the notation, because it can implicitly be recovered from V by $p = \pi(V)$, where $\pi \colon TM \to M$ is the natural projection.)

Parallel Translation

One more construction involving covariant differentiation along curves that will be useful later is parallel translation.

Let M be a manifold with a linear connection ∇. A vector field V along a curve γ is said to be *parallel along* γ with respect to ∇ if $D_tV \equiv 0$. Thus a geodesic can be characterized as a curve whose velocity vector field is parallel along the curve. A vector field V on M is said to be *parallel* if it is parallel along every curve; it is easy to check that V is parallel if and only if its total covariant derivative ∇V vanishes identically.

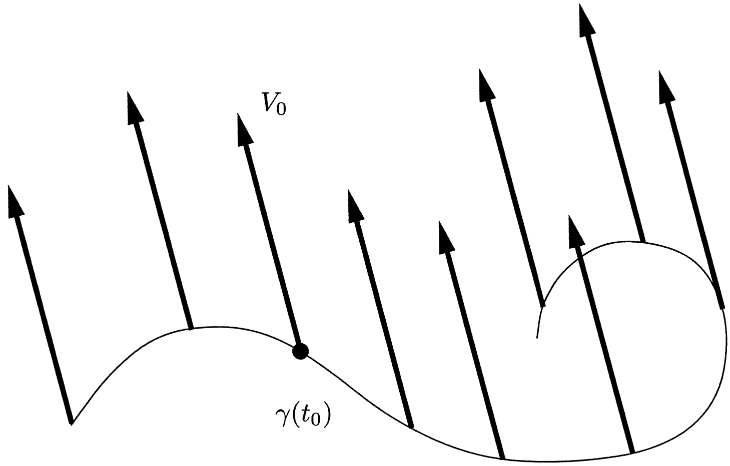


FIGURE 4.7. Parallel translate of V_0 along γ.

Exercise 4.9. Let $\gamma\colon I \to \mathbf{R}^n$ be any curve. Show that a vector field V along γ is parallel with respect to the Euclidean connection if and only if its components are constants.

The fundamental fact about parallel vector fields is that any tangent vector at any point on a curve can be uniquely extended to a parallel vector field along the entire curve.

Theorem 4.11. (Parallel Translation) *Given a curve $\gamma\colon I \to M$, $t_0 \in I$, and a vector $V_0 \in T_{\gamma(t_0)}M$, there exists a unique parallel vector field V along γ such that $V(t_0) = V_0$.*

The vector field asserted to exist in Theorem 4.11 is called the *parallel translate* of V_0 along γ (Figure 4.7). The proof of the theorem will use the following basic fact about ordinary differential equations: it says that, although in general we can only guarantee that solutions to ODEs exist for a short time, solutions to *linear* equations always exist for all time.

Theorem 4.12. (Existence and Uniqueness for Linear ODEs) *Let $I \subset \mathbf{R}$ be an interval, and for $1 \le j,k \le n$ let $A_j^k\colon I \to \mathbf{R}$ be arbitrary smooth functions. The linear initial-value problem*

$$\begin{aligned} \dot V^k(t) &= A_j^k(t)V^j(t), \\ V^k(t_0) &= B^k \end{aligned} \tag{4.12}$$

has a unique solution on all of I for any $t_0 \in I$ and any initial vector $(B^1,\dots,B^n) \in \mathbf{R}^n$.

Exercise 4.10. Prove the following *Escape Lemma:* Let Y be a vector field on a manifold M, and let $\gamma\colon (\alpha, \beta) \to M$ be an integral curve of Y. If $\beta < \infty$ and the image of γ is contained in some compact subset $K \subset M$, then γ extends to an integral curve on $(\alpha, \beta+\varepsilon)$ for some $\varepsilon > 0$. (See [Boo86, Lemma IV.5.1].)

Exercise 4.11. Prove Theorem 4.12, as follows. Consider the vector field Y on $I \times \mathbf{R}^n$ given by

$$\begin{aligned} Y^0(x^0, \ldots, x^n) &= 1, \\ Y^k(x^0, \ldots, x^n) &= A^k_j(x^0)x^j, \quad k = 1, \ldots, n. \end{aligned}$$

(a) Show that any solution to (4.12) is the projection to $\mathbf{R}^n$ of an integral curve of Y.

(b) For any compact subinterval $K \subset I$, show there exists a positive constant C such that every solution $V(t) = (V^1(t), \ldots, V^n(t))$ to (4.12) on K satisfies

$$\frac{d}{dt}(e^{-Ct}|V(t)|^2) \le 0.$$

(Here $|V(t)|$ is just the Euclidean norm.)

(c) If an integral curve of Y is defined only on some proper subinterval of I, use Exercise 4.10 above to derive a contradiction.

Proof of Theorem 4.11. First suppose $\gamma(I)$ is contained in a single coordinate chart. Then, using formula (4.10), V is parallel along γ if and only if

$$\dot V^k(t) = -V^j(t)\dot\gamma^i(t)\Gamma^k_{ij}(\gamma(t)), \qquad k = 1, \ldots, n. \tag{4.13}$$

This is a linear system of ODEs for $(V^1(t), \ldots, V^n(t))$. Thus Theorem 4.12 guarantees the existence and uniqueness of a solution on all of I with any initial condition $V(t_0) = V_0$.

Now suppose $\gamma(I)$ is not covered by a single chart. Let β denote the supremum of all $b > t_0$ for which there is a unique parallel translate on $[t_0, b]$. Clearly $\beta > t_0$, since for b close enough to t_0, $\gamma[t_0, b]$ is contained in a single chart and the above argument applies. Then a unique parallel translate V exists on $[t_0, \beta)$ (Figure 4.8). If $\beta \in I$, choose coordinates on an open set containing $\gamma(\beta-\delta, \beta+\delta)$ for some positive δ. (As usual, we assume γ has been extended to an open interval if necessary.) Then there exists a unique parallel vector field $\widetilde V$ on $(\beta-\delta, \beta+\delta)$ satisfying the initial condition $\widetilde V(\beta-\delta/2) = V(\beta-\delta/2)$. By uniqueness, $V = \widetilde V$ on their common domain, and therefore $\widetilde V$ is an extension of V past β, which is a contradiction. □

We conclude this chapter with an important remark. If $\gamma\colon I \to M$ is a curve and $t_0, t_1 \in I$, parallel translation defines an operator

$$P_{t_0 t_1}\colon T_{\gamma(t_0)}M \to T_{\gamma(t_1)}M \tag{4.14}$$

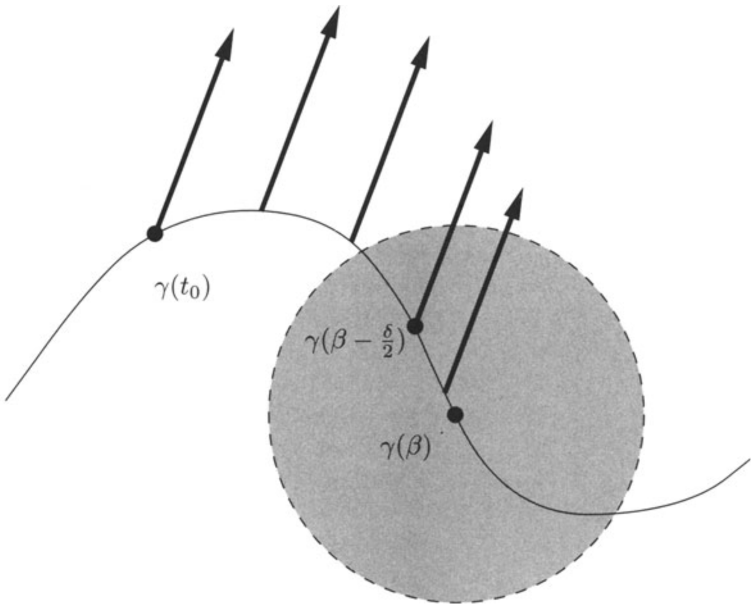


FIGURE 4.8. Existence and uniqueness of parallel translates.

by setting $P_{t_0 t_1} V_0 = V(t_1)$, where V is the parallel translate of V_0 along γ. It is easy to check that this is a linear isomorphism between $T_{\gamma(t_0)}M$ and $T_{\gamma(t_1)}M$ (because the equation of parallelism is linear). The next exercise shows that covariant differentiation along γ can be recovered from this operator. This is the sense in which a connection "connects" nearby tangent spaces.

Exercise 4.12. Let ∇ be a linear connection on M. Show that covariant differentiation along a curve γ can be recovered from parallel translation, by the following formula:

$$D_t V(t_0) = \lim_{t \to t_0} \frac{P_{t_0 t}^{-1} V(t) - V(t_0)}{t - t_0}.$$

[Hint: Use a parallel frame along γ.]

Problems

4-1. Let ∇ be a connection on M. Suppose we are given two local frames $\{E_i\}$ and $\{\widetilde{E}_j\}$ on an open subset $U \subset M$, related by $\widetilde{E}_i = A_i^j E_j$ for some matrix of functions (A_i^j). Let Γ_{ij}^k and $\widetilde{\Gamma}_{ij}^k$ denote the Christoffel symbols of ∇ with respect to these two frames. Compute a transformation law expressing Γ_{ij}^k in terms of $\widetilde{\Gamma}_{ij}^k$ and A_i^j.

4-2. Let ∇ be a linear connection on M, and define a map $\tau\colon \mathfrak{T}(M) \times \mathfrak{T}(M) \to \mathfrak{T}(M)$ by

$$\tau(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y].$$

(a) Show that τ is a $\binom{2}{1}$-tensor field, called the *torsion tensor* of ∇.

(b) We say ∇ is *symmetric* if its torsion vanishes identically. Show that ∇ is symmetric if and only if its Christoffel symbols with respect to any coordinate frame are symmetric: $\Gamma_{ij}^k = \Gamma_{ji}^k$. [*Warning:* They might not be symmetric with respect to other frames.]

(c) Show that ∇ is symmetric if and only if the covariant Hessian $\nabla^2 u$ of any smooth function $u \in C^\infty(M)$ is a symmetric 2-tensor field.

(d) Show that the Euclidean connection $\overline{\nabla}$ on $\mathbf{R}^n$ is symmetric.

4-3. In your study of differentiable manifolds, you have already seen another way of taking "directional derivatives of vector fields," the Lie derivative $\mathcal{L}_X Y$.

(a) Show that the map $\mathcal{L}\colon \mathfrak{T}(M) \times \mathfrak{T}(M) \to \mathfrak{T}(M)$ is not a connection.

(b) Show that there is a vector field on $\mathbf{R}^2$ that vanishes along the x^1-axis, but whose Lie derivative with respect to ∂_1 does not vanish on the x^1-axis. [This shows that Lie differentiation does not give a well-defined way to take directional derivatives of vector fields along curves.]

4-4. (a) If ∇^0 and ∇^1 are any two linear connections on M, show that the difference between them defines a $\binom{2}{1}$-tensor field A by

$$A(X, Y) = \nabla_X^1 Y - \nabla_X^0 Y,$$

called the *difference tensor.* Thus, if ∇^0 is any linear connection on M, the set of all linear connections is precisely $\{\nabla^0 + A : A \in \mathfrak{T}_1^2(M)\}$.

(b) Show that ∇^0 and ∇^1 determine the same geodesics if and only if their difference tensor is antisymmetric, i.e., $A(X, Y) = -A(Y, X)$.

(c) Show that ∇^0 and ∇^1 have the same torsion tensor (Problem 4-2) if and only if their difference tensor is symmetric, i.e., $A(X,Y) = A(Y,X)$.

4-5. Let ∇ be a linear connection on M, let $\{E_i\}$ be a local frame on some open subset $U \subset M$, and let $\{\varphi^i\}$ be the dual coframe.

(a) Show that there is a uniquely determined matrix of 1-forms $\omega_i{}^j$ on U, called the *connection 1-forms* for this frame, such that

$$\nabla_X E_i = \omega_i{}^j(X) E_j$$

for all $X \in TM$.

(b) Prove *Cartan's first structure equation:*

$$d\varphi^j = \varphi^i \wedge \omega_i{}^j + \tau^j,$$

where $\{\tau^1, \ldots, \tau^n\}$ are the *torsion 2-forms,* defined in terms of the torsion tensor τ (Problem 4-2) and the frame $\{E_i\}$ by

$$\tau(X,Y) = \tau^j(X,Y) E_j.$$

5

Riemannian Geodesics

If we are to use geodesics and covariant derivatives as tools for studying Riemannian geometry, it is evident that we need a way to single out a particular connection on a Riemannian manifold that reflects the properties of the metric. In this chapter, guided by the example of an embedded submanifold of $\mathbf{R}^n$, we describe two properties that determine a unique connection on any Riemannian manifold. The first property, compatibility with the metric, is easy to motivate and understand. The second, symmetry, is a bit more mysterious.

After defining the Riemannian connection and its geodesics, we investigate the exponential map, which conveniently encodes the collective behavior of geodesics and allows us to study the way they change as the initial point and initial vector vary. Having established the properties of this map, we introduce normal neighborhoods and Riemannian normal coordinates. Finally, we return to our model Riemannian manifolds and determine their geodesics.

The Riemannian Connection

We are going to show that on each Riemannian manifold there is a natural connection that is particularly suited to computations in Riemannian geometry. Since we get most of our intuition about Riemannian manifolds from studying submanifolds of $\mathbf{R}^n$ with the induced metric, let's start by examining that case. As a guiding principle, consider the idea mentioned

in the beginning of Chapter 4: A geodesic in a submanifold of $\mathbf{R}^n$ should be "as straight as possible," which we take to mean that its acceleration vector field should have zero tangential projection onto TM.

To express this in the language of connections, let $M \subset \mathbf{R}^n$ be an embedded submanifold. Any vector field on M can be extended to a smooth vector field on $\mathbf{R}^n$ by the result of Exercise 2.3(b). Define a map

$$\nabla^\top : \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$$

by setting

$$\nabla^\top_X Y := \pi^\top(\overline{\nabla}_X Y),$$

where X and Y are extended arbitrarily to $\mathbf{R}^n$, $\overline{\nabla}$ is the Euclidean connection (4.4) on $\mathbf{R}^n$, and for any point $p \in M$, $\pi^\top : T_p\mathbf{R}^n \to T_pM$ is the orthogonal projection. As the next lemma shows, this turns out to be a linear connection on M, called the *tangential connection.*

Lemma 5.1. *The operator $\nabla^\top$ is well defined, and is a connection on M.*

Proof. Since the value of $\overline{\nabla}_X Y$ at a point $p \in M$ depends only on X_p, $\nabla^\top_X Y$ is clearly independent of the choice of vector field extending X. On the other hand, because of the result of Exercise 4.7, the value of $\overline{\nabla}_X Y$ at p depends only on the values of Y along a curve whose initial tangent vector is X_p; taking the curve to lie entirely in M shows that $\nabla^\top_X Y$ depends only on the original vector field $Y \in \mathcal{T}(M)$. Thus $\nabla^\top$ is well defined. Smoothness follows easily by expressing $\overline{\nabla}_X Y$ in terms of an adapted orthonormal frame as in Problem 3-1.

It is obvious from the definition that $\nabla^\top_X Y$ is linear over $C^\infty(M)$ in X and over $\mathbf{R}$ in Y, so to show that it is a connection, only the product rule needs checking. Let $f \in C^\infty(M)$ be extended arbitrarily to $\mathbf{R}^n$. Evaluating along M, we get

$$\begin{aligned}
\nabla^\top_X(fY) &= \pi^\top(\overline{\nabla}_X(fY)) \\
&= (Xf)\pi^\top Y + f\pi^\top(\overline{\nabla}_X Y) \\
&= (Xf)Y + f\nabla^\top_X Y.
\end{aligned}$$

Thus $\nabla^\top$ is a connection. □

There is a celebrated (and hard) theorem of John Nash [Nas56] that says any Riemannian metric on any manifold can be realized as the induced metric of some embedding in a Euclidean space. Thus, in a certain sense, one would lose no generality by studying only submanifolds of $\mathbf{R}^n$ with their induced metrics, for which the tangential connection would suffice. However, when one is trying to understand *intrinsic* properties of a Riemannian manifold, an embedding introduces a great deal of extraneous

information, and in some cases actually makes it harder to discern which geometric properties depend only on the metric. Our task in this chapter is to distinguish some important properties of the tangential connection that make sense for connections on an abstract Riemannian manifold, and to use them to single out a unique connection in the abstract case.

The Euclidean connection on $\mathbf{R}^n$ has one very nice property with respect to the Euclidean metric: it satisfies the product rule

$$\overline{\nabla}_X \langle Y, Z\rangle = \langle \overline{\nabla}_X Y, Z\rangle + \langle Y, \overline{\nabla}_X Z\rangle,$$

as you can verify easily by computing in terms of the standard basis. It is almost immediate that the tangential connection has the same property, if we now interpret all the vector fields as being tangent to M and interpret the inner products as being taken with respect to the induced metric on M (see Exercise 5.2 below).

This property makes sense on an abstract Riemannian manifold, and seems so natural and desirable that it has a name. Let g be a Riemannian (or pseudo-Riemannian) metric on a manifold M. A linear connection ∇ is said to be *compatible with* g if it satisfies the following product rule for all vector fields X, Y, Z.

$$\nabla_X \langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle .$$

Lemma 5.2. *The following conditions are equivalent for a linear connection ∇ on a Riemannian manifold:*

(*a*) *∇ is compatible with g.*

(*b*) *$\nabla g \equiv 0$.*

(*c*) *If V, W are vector fields along any curve γ,*

$$\frac{d}{dt}\langle V, W\rangle = \langle D_t V, W\rangle + \langle V, D_t W\rangle .$$

(*d*) *If V, W are parallel vector fields along a curve γ, then $\langle V, W\rangle$ is constant.*

(*e*) *Parallel translation $P_{t_0 t_1} : T_{\gamma(t_0)}M \to T_{\gamma(t_1)}M$ is an isometry for each t_0, t_1 (Figure 5.1).*

Exercise 5.1. Prove Lemma 5.2.

Exercise 5.2. Prove that the tangential connection on any embedded submanifold of $\mathbf{R}^n$ is compatible with the induced Riemannian metric.

It turns out that requiring a connection to be compatible with the metric is not enough to determine a unique connection, so we turn to another key

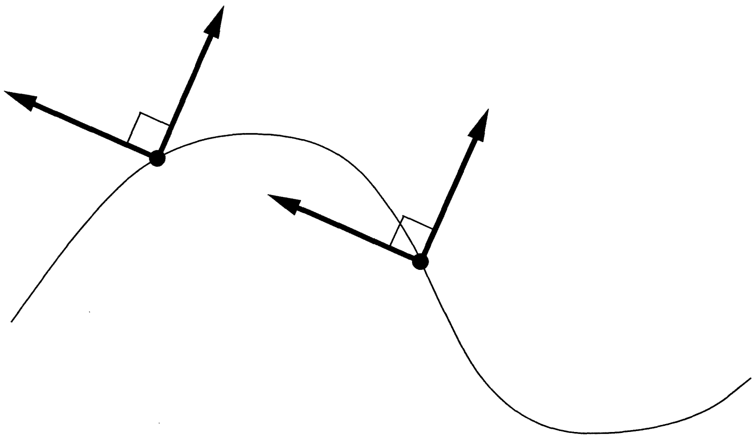


FIGURE 5.1. Parallel translation is an isometry.

property of the tangential connection. This involves the *torsion tensor* of the connection (see Problem 4-2), which is the $\binom{2}{1}$-tensor field $\tau\colon \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$ defined by

$$\tau(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y].$$

A linear connection ∇ is said to be *symmetric* if its torsion vanishes identically, that is, if

$$\nabla_X Y - \nabla_Y X \equiv [X,Y].$$

Lemma 5.3. *The tangential connection on an embedded submanifold $M \subset \mathbf{R}^n$ is symmetric.*

Exercise 5.3. Prove Lemma 5.3. [Hint: If X and Y are vector fields on $\mathbf{R}^n$ that are tangent to M at points of M, so is $[X,Y]$ by Exercise 2.3.]

Theorem 5.4. (Fundamental Lemma of Riemannian Geometry) *Let (M,g) be a Riemannian (or pseudo-Riemannian) manifold. There exists a unique linear connection ∇ on M that is compatible with g and symmetric.*

This connection is called the *Riemannian connection* or the *Levi–Civita connection* of g.

Proof. We prove uniqueness first, by deriving a formula for ∇. Suppose, therefore, that ∇ is such a connection, and let $X, Y, Z \in \mathcal{T}(M)$ be arbitrary vector fields. Writing the compatibility equation three times with X, Y, Z

cyclically permuted, we obtain

$$\begin{aligned} X\langle Y,Z\rangle &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ Y\langle Z,X\rangle &= \langle \nabla_Y Z, X\rangle + \langle Z, \nabla_Y X\rangle \\ Z\langle X,Y\rangle &= \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle. \end{aligned}$$

Using the symmetry condition on the last term in each line, this can be rewritten as

$$\begin{aligned} X\langle Y,Z\rangle &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_Z X\rangle + \langle Y, [X,Z]\rangle \\ Y\langle Z,X\rangle &= \langle \nabla_Y Z, X\rangle + \langle Z, \nabla_X Y\rangle + \langle Z, [Y,X]\rangle \\ Z\langle X,Y\rangle &= \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Y Z\rangle + \langle X, [Z,Y]\rangle. \end{aligned}$$

Adding the first two of these equations and subtracting the third, we obtain

$$\begin{aligned} X\langle Y,Z\rangle + Y\langle Z,X\rangle - Z\langle X,Y\rangle &= \\ &2\langle \nabla_X Y, Z\rangle + \langle Y, [X,Z]\rangle + \langle Z, [Y,X]\rangle - \langle X, [Z,Y]\rangle. \end{aligned}$$

Finally, solving for $\langle \nabla_X Y, Z\rangle$, we get

$$\begin{aligned} \langle \nabla_X Y, Z\rangle = \frac{1}{2}\big(&X\langle Y,Z\rangle + Y\langle Z,X\rangle - Z\langle X,Y\rangle \\ &- \langle Y,[X,Z]\rangle - \langle Z,[Y,X]\rangle + \langle X,[Z,Y]\rangle\big). \end{aligned} \tag{5.1}$$

Now suppose ∇^1 and ∇^2 are two connections that are symmetric and compatible with g. Since the right-hand side of (5.1) does not depend on the connection, it follows that $\langle \nabla^1_X Y - \nabla^2_X Y, Z\rangle = 0$ for all X, Y, Z. This can only happen if $\nabla^1_X Y = \nabla^2_X Y$ for all X and Y, so $\nabla^1 = \nabla^2$.

To prove existence, we use (5.1), or rather a coordinate version of it. It suffices to prove that such a connection exists in each coordinate chart, for then uniqueness ensures that the connections constructed in different charts agree where they overlap.

Let $(U, (x^i))$ be any local coordinate chart. Applying (5.1) to the coordinate vector fields, whose Lie brackets are zero, we obtain

$$\langle \nabla_{\partial_i} \partial_j, \partial_l\rangle = \frac{1}{2}\left(\partial_i \langle \partial_j, \partial_l\rangle + \partial_j \langle \partial_l, \partial_i\rangle - \partial_l \langle \partial_i, \partial_j\rangle\right). \tag{5.2}$$

Recall the definitions of the metric coefficients and the Christoffel symbols:

$$g_{ij} = \langle \partial_i, \partial_j\rangle, \qquad \nabla_{\partial_i}\partial_j = \Gamma^m_{ij}\partial_m.$$

Inserting these into (5.2) yields

$$\Gamma^m_{ij} g_{ml} = \frac{1}{2}\left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\right). \tag{5.3}$$

Finally, multiplying both sides by the inverse matrix g^{lk} and noting that $g_{ml}g^{lk} = \delta^k_m$, we get

$$\Gamma^k_{ij} = \frac{1}{2} g^{kl} \left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij} \right). \tag{5.4}$$

This formula certainly defines a connection in each chart, and it is evident from the formula that $\Gamma^k_{ij} = \Gamma^k_{ji}$, so the connection is symmetric by Problem 4-2(b). Thus only compatibility with the metric needs to be checked. By Lemma 5.2, it suffices to show that $\nabla g = 0$. In terms of a coordinate frame, the components of ∇g (see Lemma 4.8) are

$$g_{ij;k} = \partial_k g_{ij} - \Gamma^l_{ki} g_{lj} - \Gamma^l_{kj} g_{il}.$$

Using (5.3) twice, we conclude

$$\begin{aligned} \Gamma^l_{ki} g_{lj} + \Gamma^l_{kj} g_{il} &= \frac{1}{2} \left(\partial_k g_{ij} + \partial_i g_{kj} - \partial_j g_{ki} \right) + \frac{1}{2} \left(\partial_k g_{ji} + \partial_j g_{ki} - \partial_i g_{kj} \right) \\ &= \partial_k g_{ij}, \end{aligned}$$

which shows that $g_{ij;k} = 0$. □

A bonus of this proof is that it gives us an explicit formula (5.4) for computing the Christoffel symbols of the Riemannian connection in any coordinate chart.

On any Riemannian manifold, we will always use the Riemannian connection from now on without further comment. Geodesics with respect to this connection are called *Riemannian geodesics,* or simply *geodesics,* as long as there is no risk of confusion.

One immediate consequence of the definitions is the following lemma. If γ is a curve in a Riemannian manifold, the *speed* of γ at any time t is the length of its velocity vector $|\dot\gamma(t)|$. We say γ is *constant speed* if $|\dot\gamma(t)|$ is independent of t, and *unit speed* if the speed is identically equal to 1.

Lemma 5.5. *All Riemannian geodesics are constant speed curves.*

Proof. Let γ be a Riemannian geodesic. Since $\dot\gamma$ is parallel along γ, its length $|\dot\gamma| = \langle \dot\gamma, \dot\gamma \rangle^{1/2}$ is constant by Lemma 5.2(d). □

Another consequence of the definition is that, because they are defined in coordinate-invariant terms, Riemannian connections behave well with respect to isometries.

Proposition 5.6. (Naturality of the Riemannian Connection) *Suppose $\varphi\colon (M, g) \to (\widetilde M, \tilde g)$ is an isometry.*

(*a*) φ *takes the Riemannian connection* ∇ *of* g *to the Riemannian connection* $\widetilde{\nabla}$ *of* $\tilde{g}$, *in the sense that*

$$\varphi_*(\nabla_X Y) = \widetilde{\nabla}_{\varphi_* X}(\varphi_* Y).$$

(*b*) *If* γ *is a curve in* M *and* V *is a vector field along* γ, *then*

$$\varphi_* D_t V = \widetilde{D}_t(\varphi_* V).$$

(*c*) φ *takes geodesics to geodesics: if* γ *is the geodesic in* M *with initial point* p *and initial velocity* V, *then* $\varphi \circ \gamma$ *is the geodesic in* $\widetilde{M}$ *with initial point* $\varphi(p)$ *and initial velocity* $\varphi_* V$.

Exercise 5.4. Prove Proposition 5.6 as follows. For part (a), define a map

$$\varphi^* \widetilde{\nabla} \colon \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$$

by

$$(\varphi^* \widetilde{\nabla})_X Y = \varphi_*^{-1}(\widetilde{\nabla}_{\varphi_* X}(\varphi_* Y)).$$

Show that $\varphi^* \widetilde{\nabla}$ is a connection on M (called the *pullback connection*), and that it is symmetric and compatible with g; therefore $\varphi^* \widetilde{\nabla} = \nabla$ by uniqueness of the Riemannian connection. You will have to unwind the definition of the push-forward of a vector field very carefully. For part (b), define an operator $\varphi^* \widetilde{D}_t \colon \mathcal{T}(\gamma) \to \mathcal{T}(\gamma)$ by a similar formula and show that it is equal to D_t.

At this point, it is probably not clear why symmetry is a condition that one would want a connection on a Riemannian manifold to satisfy. One important feature that recommends it is the fact that the geodesics of the Riemannian connection are locally minimizing (see Theorem 6.12 in the next chapter). Indeed, the symmetry of the connection plays a decisive role in the proof. However, this consideration alone does not force the connection to be symmetric, as you will show in Problem 6-1. A deeper reason for singling out the symmetry condition is the fact that it is natural, in the sense of Proposition 5.6. Moreover, since the tangential connection on an embedded submanifold of $\mathbf{R}^n$ is symmetric and compatible with the metric, the Riemannian connection must coincide with the tangential connection in that case. The real reason why this connection has been anointed as "the" Riemannian connection is this: symmetry and compatibility are invariantly-defined and natural properties that force the connection to coincide with the tangential connection whenever M is realized as a submanifold of $\mathbf{R}^n$ with the induced metric (which the Nash embedding theorem [Nas56] guarantees is always possible).

The Exponential Map

To further our understanding of Riemannian geodesics, we need to study their collective behavior, and in particular, to address the following question: How do geodesics change if we vary the initial point and/or the initial vector? The dependence of geodesics on the initial data is encoded in a map from the tangent bundle into the manifold, called the exponential map, whose properties are fundamental to the further study of Riemannian geometry.

In Chapter 4 we saw that any initial point $p \in M$ and any initial velocity vector $V \in T_pM$ determine a unique maximal geodesic γ_V (see the remark after Theorem 4.10). This implicitly defines a map from the tangent bundle to the set of geodesics in M. More importantly, it allows us to define a map from (a subset of) the tangent bundle to M itself, by sending the vector V to the point obtained by following γ_V for time 1.

We note in passing that the results of this section apply with only minor changes to pseudo-Riemannian metrics, or indeed to any linear connection.

Definition and Basic Properties

To be precise, define a subset $\mathcal{E}$ of TM, the *domain of the exponential map*, by

$$\mathcal{E} := \{V \in TM : \gamma_V \text{ is defined on an interval containing } [0,1]\},$$

and then define the *exponential map* $\exp\colon \mathcal{E} \to M$ by

$$\exp(V) = \gamma_V(1).$$

For each $p \in M$, the *restricted exponential map* $\exp_p$ is the restriction of $\exp$ to the set $\mathcal{E}_p := \mathcal{E} \cap T_pM$. (Some authors use the notation "Exp" to distinguish the Riemannian exponential map from the exponential map of a Lie group, but we follow the more common convention of writing both maps with a lowercase "e," since there will be very little opportunity to confuse the two.)

Proposition 5.7. (Properties of the Exponential Map)

(*a*) *$\mathcal{E}$ is an open subset of TM containing the zero section, and each set $\mathcal{E}_p$ is star-shaped with respect to 0.*

(*b*) *For each $V \in TM$, the geodesic γ_V is given by*

$$\gamma_V(t) = \exp(tV)$$

for all t such that either side is defined.

(*c*) *The exponential map is smooth.*

(Recall that a subset S of a vector space is star-shaped with respect to $x \in S$ if whenever $y \in S$, so is the line segment from x to y.)

Before proving the proposition, it is useful to prove the following simple rescaling property of geodesics.

Lemma 5.8. (Rescaling Lemma) *For any $V \in TM$ and $c, t \in R$,*

$$\gamma_{cV}(t) = \gamma_V(ct), \tag{5.5}$$

whenever either side is defined.

Proof. It suffices to show that $\gamma_{cV}(t)$ exists and (5.5) holds whenever the right-hand side is defined, for then the converse statement follows by replacing V by cV, t by ct, and c by $1/c$.

Suppose the domain of γ_V is the open interval $I \subset \mathbf{R}$. For simplicity, write $\gamma = \gamma_V$, and define a new curve $\tilde{\gamma}$ by $\tilde{\gamma}(t) = \gamma(ct)$, defined on $c^{-1}I := \{t : ct \in I\}$. We will show that $\tilde{\gamma}$ is a geodesic with initial point p and initial velocity cV; it then follows by uniqueness that it must be equal to γ_{cV}.

It is immediate from the definition that $\tilde{\gamma}(0) = \gamma(0) = p$. Writing $\gamma(t) = (\gamma^1(t), \ldots, \gamma^n(t))$ in any local coordinates, the chain rule gives

$$\begin{aligned}\dot{\tilde{\gamma}}^i(t) &= \frac{d}{dt}\gamma^i(ct) \\ &= c\dot{\gamma}^i(ct).\end{aligned}$$

In particular, it follows that $\dot{\tilde{\gamma}}(0) = c\dot{\gamma}(0) = cV$.

Now let D_t and $\widetilde{D}_t$ denote the covariant differentiation operators along γ and $\tilde{\gamma}$, respectively. Using the chain rule again in coordinates,

$$\begin{aligned}\widetilde{D}_t\dot{\tilde{\gamma}}(t) &= \left(\frac{d}{dt}\dot{\tilde{\gamma}}^k(t) + \Gamma^k_{ij}(\tilde{\gamma}(t))\dot{\tilde{\gamma}}^i(t)\dot{\tilde{\gamma}}^j(t)\right)\partial_k \\ &= \left(c^2\ddot{\gamma}^k(ct) + c^2\Gamma^k_{ij}(\gamma(ct))\dot{\gamma}^i(ct)\dot{\gamma}^j(ct)\right)\partial_k \\ &= c^2 D_t\dot{\gamma}(ct) = 0.\end{aligned}$$

Thus $\tilde{\gamma}$ is a geodesic, and so $\tilde{\gamma} = \gamma_{cV}$ as claimed. □

Proof of Proposition 5.7. The rescaling lemma with $t = 1$ says precisely that $\exp(cV) = \gamma_{cV}(1) = \gamma_V(c)$ whenever either side is defined; this is (b). Moreover, if $V \in \mathcal{E}_p$, by definition γ_V is defined at least on $[0,1]$. Thus for $0 \le t \le 1$, the rescaling lemma says that

$$\exp(tV) = \gamma_{tV}(1) = \gamma_V(t)$$

is defined. This shows that $\mathcal{E}_p$ is star-shaped.

It remains to show that $\mathcal{E}$ is open and exp is smooth. To do so, we revisit the proof of the existence and uniqueness theorem for geodesics (Theorem

4.10) and reformulate it in a more invariant way. Let (x^i) be any local coordinates on an open set $U \subset M$, and let (x^i, v^i) denote the standard coordinates for $\pi^{-1}(U) \subset TM$ constructed after the proof of Lemma 2.2.

Define a vector field G on $\pi^{-1}(U)$ by

$$G_{(x,v)} = v^k \frac{\partial}{\partial x^k} - v^i v^j \Gamma^k_{ij}(x) \frac{\partial}{\partial v^k}. \tag{5.6}$$

The integral curves of G satisfy the system of ODEs

$$\begin{aligned} \dot{x}^k(t) &= v^k(t), \\ \dot{v}^k(t) &= -v^i(t) v^j(t) \Gamma^k_{ij}(x(t)). \end{aligned} \tag{5.7}$$

This is exactly the first-order system equivalent to the geodesic equation under the substitution $v^k = \dot{x}^k$, as we observed in the proof of Theorem 4.10. Stated somewhat more invariantly, the integral curves of G on $\pi^{-1}(U)$ project to geodesics under the projection $\pi\colon TM \to M$ (which in these coordinates is just $\pi(x(t), v(t)) = x(t)$); conversely, any geodesic $\gamma(t) = (x^1(t), \ldots, x^n(t))$ lifts to an integral curve of G by setting $v^i(t) = \dot{x}^i(t)$.

The importance of G stems from the fact that it actually extends to a *global* vector field on the total space of the tangent bundle TM, called the *geodesic vector field.* The key observation, to be proved below, is that for any $f \in C^\infty(TM)$, G acts on f by

$$Gf(p, V) = \left.\frac{d}{dt}\right|_{t=0} f(\gamma_V(t), \dot{\gamma}_V(t)). \tag{5.8}$$

(Here and whenever convenient, we use the notations (p, V) and V interchangeably for an element $V \in T_pM$, depending on whether we wish to emphasize the point at which V is tangent.) Since this formula is independent of coordinates, it shows that the various definitions of G given by (5.6) in different coordinate systems agree.

To prove that G satisfies (5.8), we write the components of the geodesic $\gamma_V(t)$ as $x^i(t)$ and those of its velocity vector field as $v^i(t) = \dot{x}^i(t)$. Using the chain rule and the geodesic equation in the form (5.7), the right-hand side of (5.8) becomes

$$\begin{aligned} &\left[\frac{\partial f}{\partial x^k}(x(t), v(t))\dot{x}^k(t) + \frac{\partial f}{\partial v^k}(x(t), v(t))\dot{v}^k(t)\right]_{t=0} \\ &\qquad = \frac{\partial f}{\partial x^k}(p, V)V^k - \frac{\partial f}{\partial v^k}(p, V)V^i V^j \Gamma^k_{ij}(p) \\ &\qquad = Gf(p, V). \end{aligned}$$

The standard results on global flows of vector fields [Boo86, Theorems IV.4.3 and IV.4.5] show that there is an open neighborhood $\mathcal{O}$ of $\{0\} \times TM$ in $\mathbf{R} \times TM$ and a smooth map $\theta\colon \mathcal{O} \to TM$ such that each curve $\theta_{(p,V)}(t) =$

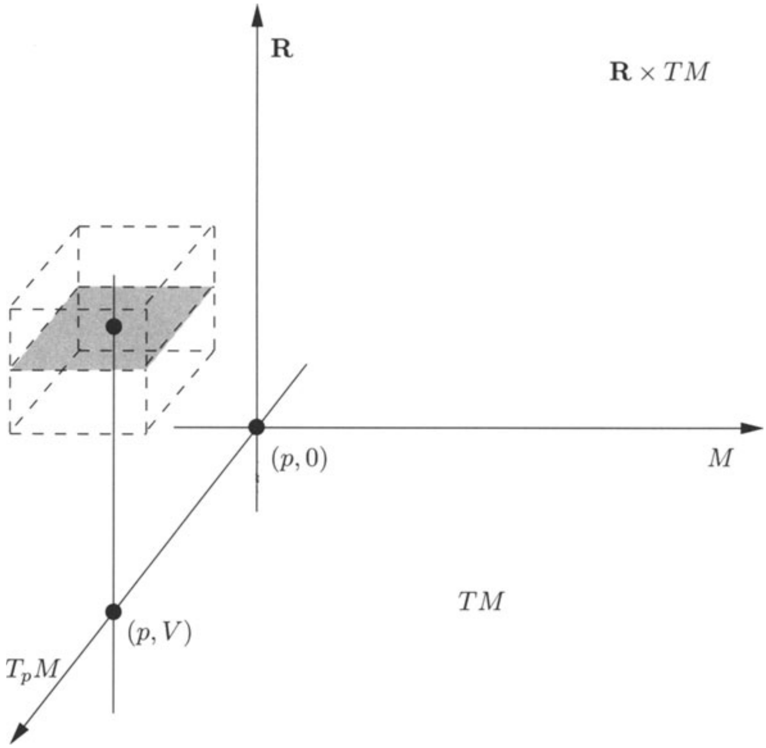


FIGURE 5.2. $\mathcal{E}$ is open.

$\theta(t, (p, V))$ is the integral curve of G starting at (p, V), defined on an open interval containing 0.

Now suppose $(p, V) \in \mathcal{E}$. This means that the geodesic γ_V is defined at least on the interval $[0, 1]$, and therefore so is the integral curve of G starting at $(p, V) \in TM$. Since $(1, (p, V)) \in \mathcal{O}$, there is an open neighborhood of $(1, (p, V))$ in $\mathbf{R} \times TM$ on which the flow of G is defined (Figure 5.2). In particular, this means there is an open neighborhood of (p, V) on which the flow exists for $t \in [0, 1]$, and therefore on which the exponential map is defined. This shows that $\mathcal{E}$ is open.

Finally, since geodesics are projections of integral curves of G, it follows that the exponential map can be expressed as

$$\exp(V) = \gamma_V(1) = \pi \circ \theta(1, (p, V))$$

wherever it is defined, and therefore exp is smooth. □

The naturality of the Riemannian connection (Proposition 5.6) and uniqueness of geodesics translate into the following important naturality property of the exponential map:

Proposition 5.9. (Naturality of the Exponential Map) *Suppose that* $\varphi\colon (M,g) \to (\widetilde{M},\tilde{g})$ *is an isometry. Then, for any* $p \in M$, *the following diagram commutes:*

$$\begin{array}{ccc} T_pM & \xrightarrow{\varphi_*} & T_{\varphi(p)}\widetilde{M} \\ {\scriptstyle \exp_p}\downarrow & & \downarrow{\scriptstyle \exp_{\varphi(p)}} \\ M & \xrightarrow[\varphi]{} & \widetilde{M} \end{array}$$

Exercise 5.5. Prove Proposition 5.9.

Normal Neighborhoods and Normal Coordinates

Recall that for any $p \in M$, the restricted exponential map $\exp_p$ maps the open subset $\mathcal{E}_p$ of the tangent space T_pM into M.

Lemma 5.10. (Normal Neighborhood Lemma) *For any* $p \in M$, *there is a neighborhood* $\mathcal{V}$ *of the origin in* T_pM *and a neighborhood* $\mathcal{U}$ *of* p *in* M *such that* $\exp_p\colon \mathcal{V} \to \mathcal{U}$ *is a diffeomorphism.*

Proof. This follows immediately from the inverse function theorem, once we show that $(\exp_p)_*$ is invertible at 0. Since T_pM is a vector space, there is a natural identification $T_0(T_pM) = T_pM$. Under this identification, we will show that $(\exp_p)_*\colon T_0(T_pM) = T_pM \to T_pM$ has a particularly simple expression: it is the identity map!

To compute $(\exp_p)_*V$ for an arbitrary vector $V \in T_pM$, we just need to choose a curve τ in T_pM starting at 0 whose initial tangent vector is V, and compute the initial tangent vector of the composite curve $\exp_p \circ \tau(t)$. An obvious such curve is $\tau(t) = tV$. Thus

$$(\exp_p)_*V = \left.\frac{d}{dt}\right|_{t=0} (\exp_p \circ \tau)(t) = \left.\frac{d}{dt}\right|_{t=0} \exp_p(tV) = \left.\frac{d}{dt}\right|_{t=0} \gamma_V(t) = V.$$

□

Any open neighborhood $\mathcal{U}$ of $p \in M$ that is the diffeomorphic image under $\exp_p$ of a star-shaped open neighborhood of $0 \in T_pM$ as in the preceding lemma is called a *normal neighborhood* of p. If $\varepsilon > 0$ is such that $\exp_p$ is a diffeomorphism on the ball $B_\varepsilon(0) \subset T_pM$ (where the radius of the ball is measured with respect to the norm defined by g), then the image set $\exp_p(B_\varepsilon(0))$ is called a *geodesic ball* in M. Also, if the closed ball $\overline{B}_\varepsilon(0)$ is contained in an open set $\mathcal{V} \subset T_pM$ on which $\exp_p$ is a diffeomorphism, then $\exp_p(\overline{B}_\varepsilon(0))$ is called a *closed geodesic ball,* and $\exp_p(\partial\overline{B}_\varepsilon(0))$ is called a *geodesic sphere.*

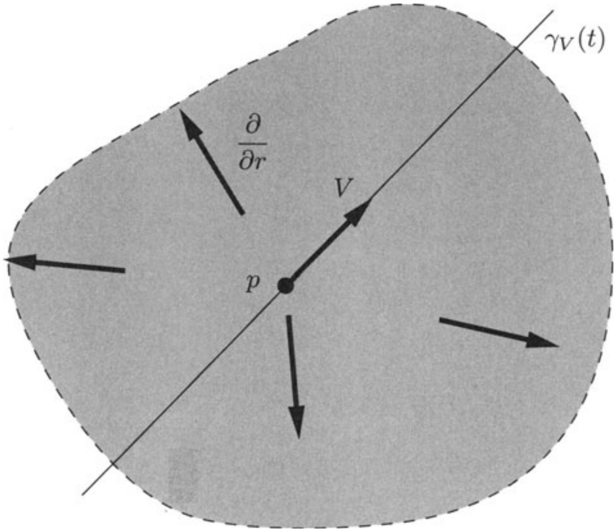


FIGURE 5.3. Riemannian normal coordinates.

An orthonormal basis $\{E_i\}$ for T_pM gives an isomorphism $E\colon \mathbf{R}^n \to T_pM$ by $E(x^1, \ldots, x^n) = x^i E_i$. If $\mathcal{U}$ is a normal neighborhood of p, we can combine this isomorphism with the exponential map to get a coordinate chart

$$\varphi := E^{-1} \circ \exp_p^{-1} \colon \mathcal{U} \to \mathbf{R}^n.$$

Any such coordinates are called (*Riemannian*) *normal coordinates* centered at p. Given $p \in M$ and a normal neighborhood $\mathcal{U}$ of p, there is a one-to-one correspondence between normal coordinate charts and orthonormal bases at p.

In any normal coordinate chart centered at p, define the *radial distance function* r by

$$r(x) := \Big(\sum_i (x^i)^2\Big)^{1/2}, \tag{5.9}$$

and the *unit radial vector field* $\partial/\partial r$ by

$$\frac{\partial}{\partial r} := \frac{x^i}{r}\frac{\partial}{\partial x^i}. \tag{5.10}$$

(See Figure 5.3.) In Euclidean space, $r(x)$ is the distance to the origin, and $\partial/\partial r$ is the unit vector field tangent to straight lines through the origin. As the next proposition shows, they also have special geometric meaning for any metric in normal coordinates. (We will strengthen these results considerably in the next chapter.)

Proposition 5.11. (Properties of Normal Coordinates) *Let $(\mathcal{U}, (x^i))$ be any normal coordinate chart centered at p.*

(*a*) *For any $V = V^i\partial_i \in T_pM$, the geodesic γ_V starting at p with initial velocity vector V is represented in normal coordinates by the radial line segment*

$$\gamma_V(t) = (tV^1, \ldots, tV^n) \tag{5.11}$$

as long as γ_V stays within $\mathcal{U}$.

(*b*) *The coordinates of p are $(0, \ldots, 0)$.*

(*c*) *The components of the metric at p are $g_{ij} = \delta_{ij}$.*

(*d*) *Any Euclidean ball $\{x : r(x) < \varepsilon\}$ contained in $\mathcal{U}$ is a geodesic ball in M.*

(*e*) *At any point $q \in \mathcal{U} - p$, $\partial/\partial r$ is the velocity vector of the unit speed geodesic from p to q, and therefore has unit length with respect to g.*

(*f*) *The first partial derivatives of g_{ij} and the Christoffel symbols vanish at p.*

Normal coordinates are a vital tool for calculations in Riemannian geometry, so you should make sure you thoroughly understand the properties expressed in the preceding proposition. The proofs are all straightforward consequences of the fact that geodesics starting at p have the simple formula (5.11) in normal coordinates. Because of this formula, the geodesics starting at p and lying in a normal neighborhood of p are called *radial geodesics.* (But be warned that geodesics that do not pass through p do not in general have a simple form in normal coordinates.)

Exercise 5.6. Prove Proposition 5.11.

For later use in studying minimizing properties of geodesics, we need the following refinement of the concept of normal neighborhoods. An open set $\mathcal{W} \subset M$ is called *uniformly normal* if there exists some $\delta > 0$ such that $\mathcal{W}$ is contained in a geodesic ball of radius δ around each of its points (Figure 5.4).

The proof of the next lemma is fairly technical, thought not really hard. You might wish to read the statement now, and come back to the proof later.

Lemma 5.12. (Uniformly Normal Neighborhood Lemma) *Given $p \in M$ and any neighborhood $\mathcal{U}$ of p, there exists a uniformly normal neighborhood $\mathcal{W}$ of p contained in $\mathcal{U}$.*

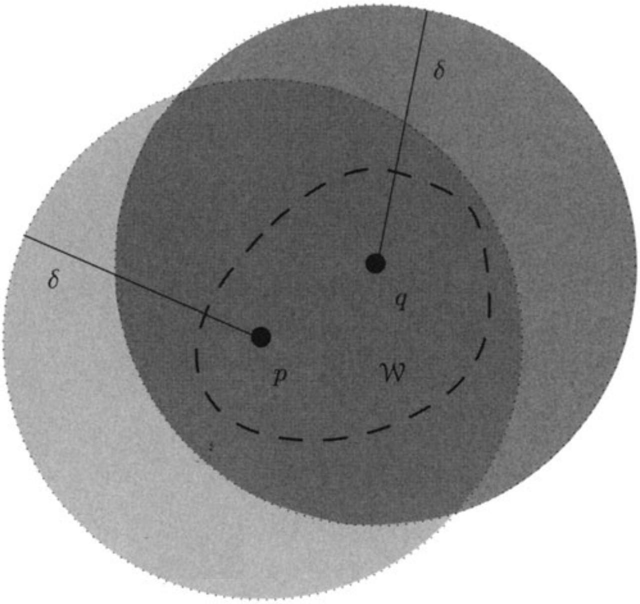


FIGURE 5.4. Uniformly normal neighborhood.

Proof. Recall that the exponential map is defined on an open subset $\mathcal{E}$ of TM. Define a new map $F\colon \mathcal{E} \to M \times M$ by

$$F(q, V) = (q, \exp_q V).$$

Choose a normal coordinate chart (x^i) for M centered at p, and let (x^i, v^i) denote the corresponding standard coordinates on TM. In these coordinates, the Jacobian matrix of F at $(p, 0)$ can be written as

$$F_* = \begin{pmatrix} \dfrac{\partial x^i}{\partial x^j} & \dfrac{\partial x^i}{\partial v^j} \\ \\ \dfrac{\partial \exp^i}{\partial x^j} & \dfrac{\partial \exp^i}{\partial v^j} \end{pmatrix} = \begin{pmatrix} Id & 0 \\ * & Id \end{pmatrix},$$

which is invertible. Thus, by the inverse function theorem, F is a diffeomorphism from some neighborhood $\mathcal{O}$ of $(p, 0)$ in TM to its image (Figure 5.5).

For any open set $\mathcal{Y} \subset M$ and any $\delta > 0$, let $\mathcal{Y}_\delta$ denote the subset of TM given by

$$\mathcal{Y}_\delta = \{(p, v) \in TM : p \in \mathcal{Y}, |v| < \delta\},$$

where as usual $|\cdot|$ denotes the norm given by g. By writing the inequality $|v| < \delta$ in any standard coordinates, it is easy to see that $\mathcal{Y}_\delta$ is open in

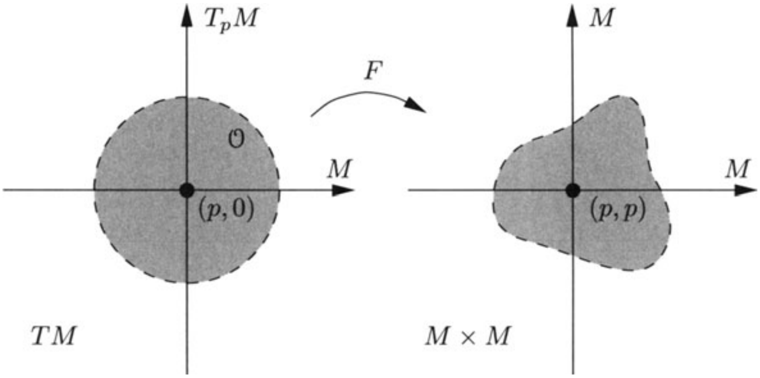


FIGURE 5.5. F is a diffeomorphism near $(p, 0)$.

the topology of TM. We will show that there is some set of the form $\mathcal{Y}_\delta$ such that $(p, 0) \in \mathcal{Y}_\delta \subset \mathcal{O}$. Since the topology on TM is generated by product open sets in local trivializations, there exists $\varepsilon > 0$ such that the set $\mathcal{X} = \{(x, v) : r(x) < 2\varepsilon,\ |v|_{\bar{g}} < 2\varepsilon\}$ is contained in $\mathcal{O}$ (Figure 5.6), where $|\cdot|_{\bar{g}}$ is the Euclidean norm in these coordinates. On the compact set $K = \{(x, v) : r(x) \leq \varepsilon,\ |v|_{\bar{g}} = \varepsilon\}$, the g-norm is continuous and nonvanishing, and therefore is bounded above and below by positive constants. Since both norms are homogeneous in the sense that $|\lambda v| = \lambda |v|$ for any positive constant λ, it follows that $c|v|_{\bar{g}} \leq |v|_g \leq C|v|_{\bar{g}}$ whenever $v \in T_xM$, $r(x) \leq \varepsilon$.

Now let $\mathcal{Y}$ be the geodesic ball $\mathcal{Y} = \{x : r(x) < \varepsilon\} \subset M$, and let $\delta = c\varepsilon$. Whenever $(x, v) \in \mathcal{Y}_\delta$, our choices guarantee that $|v|_{\bar{g}} \leq (1/c)|v|_g < \varepsilon$, so $\mathcal{Y}_\delta \subset \mathcal{X} \subset \mathcal{O}$.

Since F is a diffeomorphism on $\mathcal{Y}_\delta$ and takes $(p, 0)$ to (p, p), there is a product open set $\mathcal{W} \times \mathcal{W} \subset M \times M$ such that $(p, p) \in \mathcal{W} \times \mathcal{W} \subset F(\mathcal{Y}_\delta)$. Shrinking $\mathcal{W}$ if necessary, we may also assume that $\mathcal{W} \subset \mathcal{Y}$. We make two claims about the set $\mathcal{W}$: for any $q \in \mathcal{W}$, (1) $\exp_q$ is a diffeomorphism on $B_\delta(0) \subset T_qM$; and (2) $\mathcal{W} \subset \exp_q(B_\delta(0))$. It follows from these claims that $\mathcal{W}$ is the required uniformly normal neighborhood of p.

To prove claim (1), observe first that for each $q \in \mathcal{W}$, $\exp_q$ is at least *defined* on $B_\delta(0) \subset T_qM$; F is defined on the set $\mathcal{Y}_\delta$, so $F(q, V) = (q, \exp_q V)$ is defined whenever $|V|_g < \delta$. Because F has the form $F(q, V) = (q, \exp_q V)$, its inverse has the similar form $F^{-1}(q, y) = (q, \varphi(q, y))$ for some smooth map φ. Let's use the notation $\varphi_q(y) = \varphi(q, y)$. Then, because $F^{-1} \circ F$ is the identity on $\mathcal{Y}_\delta$, it follows that $\varphi_q \circ \exp_q$ is the identity on $B_\delta(0) \subset T_qM$ for each $q \in \mathcal{W} \subset \mathcal{Y}$. Similarly, $F \circ F^{-1} = Id$ on $F(\mathcal{Y}_\delta)$ implies that $\exp_q \circ \varphi_q$ is the identity on $\exp_q(B_\delta(0))$, so claim (1) is proved.

Finally, we turn to claim (2). Let $(q, y) \in \mathcal{W} \times \mathcal{W}$ be arbitrary. Since $\mathcal{W} \times \mathcal{W} \subset F(\mathcal{Y}_\delta)$, there is some $V \in B_\delta(0) \subset T_qM$ such that $(q, y) =$

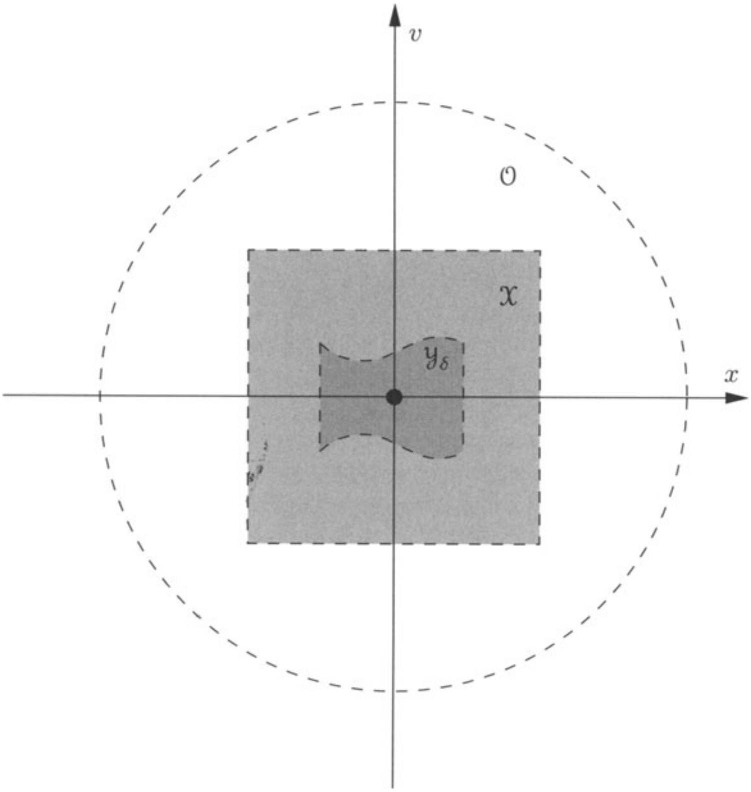


FIGURE 5.6. The sets $\mathcal{Y}_\delta \subset \mathcal{X} \subset \mathcal{O}$.

$F(q, V) = (q, \exp_q V)$. This says precisely that $y = \exp_q V$, which was to be proved. □

Geodesics of the Model Spaces

In this section we determine the geodesics of our three classes of model Riemannian manifolds defined in Chapter 3.

Euclidean Space

On $\mathbf{R}^n$ with the Euclidean metric $\bar{g}$, the metric coefficients are constants in the standard coordinate system, so it follows immediately from (5.4) that the Christoffel symbols are all zero in these coordinates. This means that the Riemannian connection on Euclidean space is exactly the Euclidean connection (4.4). Therefore, as one would expect, the Euclidean geodesics

are straight lines, and constant-coefficient vector fields are parallel (Exercises 4.8 and 4.9).

Spheres

On the 2-sphere $\mathbf{S}^2_R$ of radius R, it is not terribly difficult to compute the Christoffel symbols directly in, say, spherical coordinates, and to show that the meridians (lines of constant longitude) are geodesics, as in the following exercise.

> **Exercise 5.7.** Define spherical coordinates (θ, φ) on the subset $\mathbf{S}^2_R - \{(x, y, z) : x \leq 0, y = 0\}$ of the sphere by
>
> $$(x, y, z) = (R \sin\varphi \cos\theta, R \sin\varphi \sin\theta, R \cos\varphi), \quad -\pi < \theta < \pi, 0 < \varphi < \pi.$$
>
> (These are a special case of the coordinates for surfaces of revolution constructed in Exercise 3.3.)
>
> (a) Show that the round metric of radius R is $\mathring{g}_R = R^2 d\varphi^2 + R^2 \sin^2\varphi \, d\theta^2$ in spherical coordinates.
>
> (b) Compute the Christoffel symbols of $\mathring{g}_R$ in spherical coordinates.
>
> (c) Using the geodesic equation (4.11) in spherical coordinates, verify that each meridian $(\theta(t), \varphi(t)) = (\theta_0, t)$ is a geodesic.

As you can see from doing this exercise, even in a simple case like this, verifying the geodesic equation directly can involve a rather large number of calculations. When the metric is more complicated or the number of dimensions is high (or even when we attempt to identify arbitrary geodesics on the 2-sphere instead of just the "vertical" ones), this direct approach can become prohibitively difficult, so we must often look for other techniques to analyze geodesics.

Fortunately, the fact that the sphere is homogeneous and isotropic gives us a much easier way to determine the geodesics in all dimensions.

Proposition 5.13. *The geodesics on $\mathbf{S}^n_R$ are precisely the "great circles" (intersections of $\mathbf{S}^n_R$ with 2-planes through the origin), with constant speed parametrizations.*

Proof. First we consider a geodesic $\gamma(t) = (x^1(t), \ldots, x^{n+1}(t))$ starting at the north pole N whose initial velocity V is a multiple of ∂_1. It is intuitively evident by symmetry that this geodesic must remain along the meridian $x^2 = \cdots = x^n = 0$. To make this intuition rigorous, suppose not; that is, suppose there were a time t_0 such that $x^i(t_0) \neq 0$ for some $2 \leq i \leq n$. The linear map $\varphi\colon \mathbf{R}^{n+1} \to \mathbf{R}^{n+1}$ sending x^i to $-x^i$ and leaving the other coordinates fixed is an isometry of the sphere that fixes $N = \gamma(0)$ and $V = \dot\gamma(0)$, and therefore it takes γ to γ. But $\varphi(\gamma(t_0)) \neq \gamma(t_0)$, a contradiction (Figure 5.7).

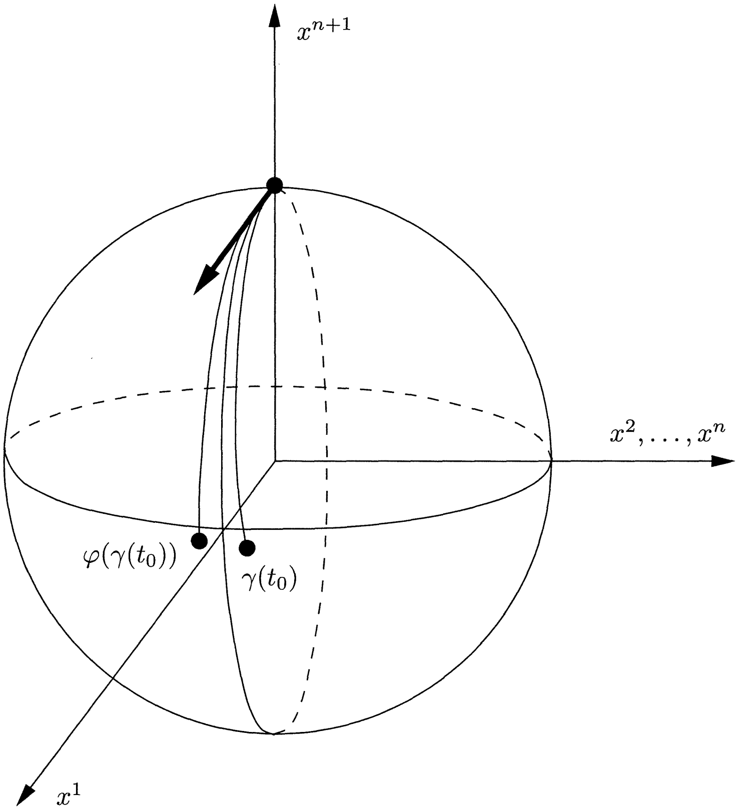


FIGURE 5.7. Identifying geodesics on S^2 by symmetry.

Since geodesics have constant speed, the geodesic with initial point N and initial velocity $c\partial_1$ must therefore be the circle where $\mathbf{S}^n_R$ intersects the (x^1, x^{n+1})-plane, with a constant speed parametrization. Since there is an orthogonal map taking any other initial point to N and any other initial vector to one of this form, and since orthogonal maps take planes through the origin to planes through the origin, it follows that the geodesics on $\mathbf{S}^n_R$ are precisely the intersections of $\mathbf{S}^n_R$ with 2-planes through the origin. □

Hyperbolic Spaces

The geodesics of $\mathbf{H}^n_R$ are easily determined using homogeneity and isotropy, as in the case of the sphere.

Proposition 5.14. *The geodesics on the hyperbolic spaces are the following curves, with constant speed parametrizations:*

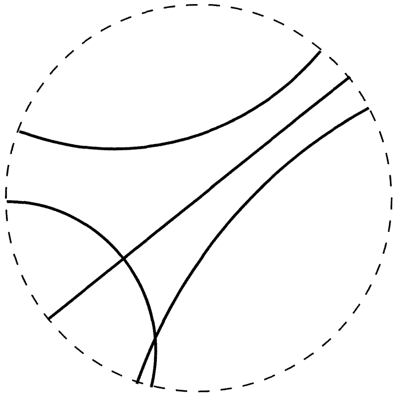


FIGURE 5.8. Geodesics of the Poincaré ball.

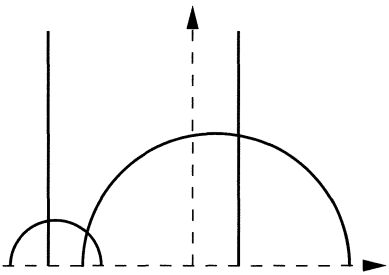


FIGURE 5.9. Geodesics of the Poincaré half-space.

HYPERBOLOID MODEL: *The "great hyperbolas," or intersections of* $\mathbf{H}_R^n$ *with* 2*-planes through the origin.*

BALL MODEL: *The line segments through the origin and the circular arcs that intersect* $\partial \mathbf{B}_R^n$ *orthogonally (Figure 5.8).*

HALF-SPACE MODEL: *The vertical half-lines and the semicircles with centers on the* $y = 0$ *hyperplane (Figure 5.9).*

Proof. We begin with the hyperboloid model. As with the sphere, the geodesic starting at N with initial tangent vector parallel to $\partial/\partial\xi^1$ must remain in the (ξ^1, τ) plane by symmetry (Figure 5.10), and therefore must be a constant speed parametrization of the hyperbola where this plane intersects $\mathbf{H}_R^n$. Since $\mathbf{H}_R^n$ is homogeneous and isotropic, and $O_+(n,1)$ takes 2-planes through the origin to 2-planes through the origin, the result follows.

For the ball model, first consider the 2-dimensional case, and recall the hyperbolic stereographic projection $\pi\colon \mathbf{H}_R^2 \to \mathbf{B}_R^2$ constructed in Chapter 3:

$$\pi(\xi,\tau) = u = \frac{R\xi}{R+\tau},$$
$$\pi^{-1}(u) = (\xi,\tau) = \left(\frac{2R^2u}{R^2 - |u|^2}, R\frac{R^2+|u|^2}{R^2-|u|^2}\right).$$

A geodesic in the hyperboloid model is the set of points on $\mathbf{H}_R^2$ that solve a linear equation $\alpha_i\xi^i + \beta\tau = 0$, with a constant speed parametrization. In the special case $\beta = 0$, this hyperbola is mapped by π to a straight line segment through the origin, as can easily be seen from the geometric definition of π. If $\beta \neq 0$, we can divide through by $-\beta$ and write the linear

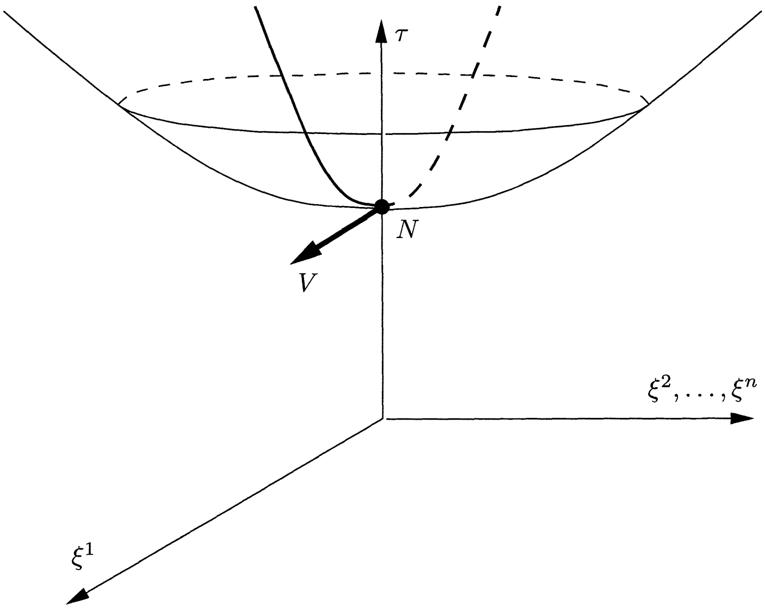


FIGURE 5.10. A great hyperbola.

equation as $\tau = \alpha_i \xi^i = \langle \alpha, \xi \rangle$ (for a different covector α). Under π^{-1}, this pulls back to the equation

$$R\frac{R^2 + |u|^2}{R^2 - |u|^2} = \frac{2R^2 \langle \alpha, u \rangle}{R^2 - |u|^2}$$

on the disk, which simplifies to

$$|u|^2 - 2R\langle \alpha, u \rangle + R^2 = 0.$$

Completing the square, we can write this as

$$|u - R\alpha|^2 = R^2(|\alpha|^2 - 1). \tag{5.12}$$

If $|\alpha|^2 \leq 1$ this locus is either empty or a point on $\partial\mathbf{B}^2_R$, so it does not define a geodesic. When $|\alpha|^2 > 1$, this is the circle with center $R\alpha$ and radius $R\sqrt{|\alpha|^2 - 1}$. At a point u_0 where the circle intersects $\partial\mathbf{B}^2_R$, the three points 0, u_0, and $R\alpha$ form a triangle with sides $|u_0| = R$, $|R\alpha|$, and $|u_0 - R\alpha|$ (Figure 5.11), which satisfy the Pythagorean identity by (5.12); therefore the circle meets $\partial\mathbf{B}^2_R$ in a right angle. By the existence and uniqueness theorem, it is easy to see that the line segments through the origin and the circular arcs that intersect $\partial\mathbf{B}^2_R$ orthogonally are all the geodesics.

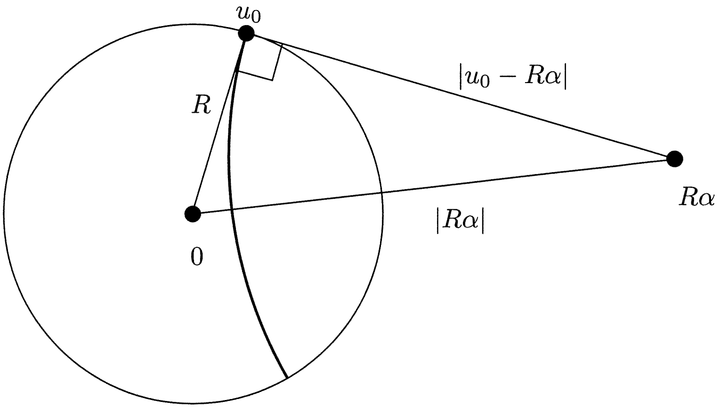


FIGURE 5.11. Geodesics are arcs of circles orthogonal to the boundary.

In the higher-dimensional case, a geodesic on $\mathbf{H}^n_R$ is determined by a 2-plane. If the 2-plane contains the point N, the corresponding geodesic on $\mathbf{B}^n_R$ is a line through the origin as before. Otherwise, we can conjugate with an orthogonal transformation in the $(\xi^1, \dots, \xi^n)$ variables (which preserves h_R) to move this 2-plane so that it lies in the (ξ^1, ξ^{n+1}, τ) subspace, and then we are in the same situation as in the 2-dimensional case.

Now consider the upper half-space model. The 2-dimensional case is easiest to analyze using complex notation. It is straightforward to check that the inverse of the complex Cayley transform (3.12) is

$$\kappa^{-1}(z) = w = iR\frac{z - iR}{z + iR}.$$

Substituting this into equation (5.12) and writing $w = u + iv$ and $\alpha = a + ib$ in place of $u = (u^1, u^2)$, $\alpha = (\alpha^1, \alpha^2)$, we get

$$R^2\frac{|z - iR|^2}{|z + iR|^2} - iR^2\bar{\alpha}\frac{z - iR}{z + iR} + iR^2\alpha\frac{\bar{z} + iR}{\bar{z} - iR} + R^2|\alpha|^2 = R^2(|\alpha|^2 - 1).$$

Multiplying through by $(z + iR)(\bar{z} - iR)/2R^2$ and simplifying,

$$(1 - b)|z|^2 - 2aRx + (b + 1)R^2 = 0.$$

This is the equation of a circle with center on the x-axis, unless $b = 1$, in which case the condition $|\alpha|^2 > 1$ forces $a \neq 0$, and then it is a straight line $x =$ constant. The other class of geodesics on the ball, line segments through the origin, can be handled similarly.

In the higher-dimensional case, we just conjugate κ with a suitable orthogonal transformation in the first $n - 1$ variables, and apply the usual symmetry arguments to show that the resulting geodesics remain in the (u^1, v)- and (x^1, y)-planes. □

Problems

5-1. Let ∇ be a linear connection on a Riemannian manifold (M, g). Show that ∇ is compatible with g if and only if the connection 1-forms $\omega_i{}^j$ (Problem 4-5) with respect to any local frame $\{E_i\}$ satisfy

$$g_{jk}\omega_i{}^k + g_{ik}\omega_j{}^k = dg_{ij}.$$

In particular, the matrix $\omega_i{}^j$ of connection 1-forms for the Riemannian connection with respect to any local orthonormal frame is skew-symmetric.

5-2. Let $M \subset \mathbf{R}^3$ be a surface of revolution, parametrized as in Exercise 3.3. It will simplify the computations if we assume that the curve γ (called a *generating curve* for the surface) is unit speed.

(a) Compute the Christoffel symbols of the induced metric in (θ, t) coordinates.

(b) Show that each "meridian" $\{\theta = \theta_0\}$ is a geodesic on M.

(c) Determine necessary and sufficient conditions for a "latitude circle" $\{t = t_0\}$ to be a geodesic.

5-3. Let $\mathbf{H}^n_R$ denote the n-dimensional hyperbolic space of radius R.

(a) Determine the unit speed parametrization of the geodesic in the hyperboloid model starting at $N = (0, \ldots, R)$ with initial tangent vector $\partial/\partial\xi^1$.

(b) Prove that each geodesic on $\mathbf{H}^n_R$ is defined for all $t \in \mathbf{R}$, and that the image of each geodesic is an entire branch of a great hyperbola.

5-4. Recall that a vector field V is said to be parallel if $\nabla V \equiv 0$.

(a) Let $p \in \mathbf{R}^n$ and $V_p \in T_p\mathbf{R}^n$. Show that V_p has a unique extension to a parallel vector field V on $\mathbf{R}^n$.

(b) Let $\mathcal{U}$ be the open subset of the unit sphere $\mathbf{S}^2$ on which spherical coordinates (θ, φ) are defined (see Exercise 5.7), and let $V = \partial/\partial\varphi$ in these coordinates. Compute $\nabla_{\frac{\partial}{\partial\theta}} V$ and $\nabla_{\frac{\partial}{\partial\varphi}} V$, and conclude that V is parallel along the equator and along each meridian $\theta = \theta_0$.

(c) Let $p = (0, \pi/2)$ in spherical coordinates. Show that V_p has no parallel extension to any neighborhood of p.

(d) Use (a) and (c) to show that no neighborhood of p is isometric to an open subset of $\mathbf{R}^2$.

5-5. Let (M, g) be a Riemannian manifold. If f is a smooth function on M such that $|\operatorname{grad} f| \equiv 1$, show that the integral curves of $\operatorname{grad} f$ are geodesics.

5-6. Let (M, g) be an oriented Riemannian manifold, and div the divergence operator defined in Problem 3-3.

(a) Show that if $X = X^i E_i$ in terms of some local frame, then $\operatorname{div} X$ can be written in terms of covariant derivatives as

$$\operatorname{div} X = X^i{}_{;i}.$$

[Hint: Show that it suffices to prove the formula at the origin in normal coordinates.]

(b) Now suppose M is a compact, oriented Riemannian manifold with boundary. Extend the integration by parts formula of Problem 3-3 as follows: If ω is any k-tensor field and η any $k+1$-tensor field,

$$\int_M \langle \nabla\omega, \eta\rangle \, dV = -\int_M \langle \omega, \operatorname{tr}_g \nabla\eta\rangle \, dV + \int_{\partial M} \langle \omega \otimes N, \eta\rangle \, d\widetilde{V},$$

where the trace is on the last two indices of $\nabla\eta$. This is often written in the suggestive but not-quite-rigorous notation

$$\int_M \omega_{i_1\dots i_k;j}\eta^{i_1\dots i_k j} \, dV$$
$$= -\int_M \omega_{i_1\dots i_k}\eta^{i_1\dots i_k j}{}_{;j} \, dV + \int_{\partial M} \omega_{i_1\dots i_k}\eta^{i_1\dots i_k j} N_j \, d\widetilde{V}.$$

5-7. If (M, g) and $(\widetilde{M}, \tilde{g})$ are Riemannian manifolds, a map $\varphi\colon M \to \widetilde{M}$ is a *local isometry* if each point $p \in M$ has a neighborhood $\mathcal{U}$ such that $\varphi|_{\mathcal{U}}$ is an isometry onto an open subset of $\widetilde{M}$. Suppose M is connected, and suppose $\varphi, \psi\colon M \to \widetilde{M}$ are local isometries such that for some point $p \in M$, $\varphi(p) = \psi(p)$ and $\varphi_* = \psi_*$ at p. Show that $\varphi \equiv \psi$.

5-8. Let $E(n)$ be the Euclidean group described in Problem 3-6.

(a) Show that $\mathcal{I}(\mathbf{R}^n) = E(n)$, $\mathcal{I}(\mathbf{H}^n_R) = O_+(n, 1)$, and $\mathcal{I}(\mathbf{S}^n_R) = O(n+1)$.

(b) Strengthen the result above by showing that if M is one of our model Riemannian manifolds ($M = \mathbf{R}^n$, $\mathbf{H}^n_R$, or $\mathbf{S}^n_R$), U, V are connected open subsets of M, and $\varphi\colon U \to V$ is an isometry, then φ is the restriction to U of an element of $\mathcal{I}(M)$.

5-9. Suppose $p\colon (\widetilde{M},\tilde{g}) \to (M,g)$ is a Riemannian submersion (Problem 3-8). A vector field on $\widetilde{M}$ is said to be *horizontal* or *vertical* if its value is in the horizontal or vertical space at each point, respectively.

(a) For any vector fields $X, Y \in \mathcal{T}(M)$, show that

$$\langle \widetilde{X}, \widetilde{Y} \rangle = p^*\langle X, Y\rangle;$$
$$[\widetilde{X}, \widetilde{Y}]^H = \widetilde{[X,Y]};$$
$$[\widetilde{X}, W] \text{ is vertical if } W \text{ is vertical.}$$

(b) Let $\widetilde{\nabla}$ and ∇ denote the Riemannian connections of $\tilde{g}$ and g, respectively. For any vector fields $X, Y \in \mathcal{T}(M)$, show that

$$\widetilde{\nabla}_{\widetilde{X}}\widetilde{Y} = \widetilde{\nabla_X Y} + \frac{1}{2}[\widetilde{X}, \widetilde{Y}]^V. \tag{5.13}$$

[Hint: Let $\widetilde{Z}$ be a horizontal lift and W be a vertical vector field on $\widetilde{M}$, and compute $\langle \nabla_{\widetilde{X}}\widetilde{Y}, \widetilde{Z}\rangle$ and $\langle \nabla_{\widetilde{X}}\widetilde{Y}, W\rangle$ using formula (5.1).]

5-10. Suppose G is a Lie group with Lie algebra $\mathfrak{g}$, and let $(X_1, \dots, X_n)$ be any basis of $\mathfrak{g}$. Define the *structure constants* c_{ij}^k by

$$[X_i, X_j] = \sum_k c_{ij}^k X_k.$$

For an arbitrary left-invariant metric g on G, compute the Christoffel symbols of the Riemannian connection (with respect to the basis $\{X_i\}$) in terms of c_{ij}^k and g_{ij}.

5-11. Let G be a Lie group and $\mathfrak{g}$ its Lie algebra, and let g be a bi-invariant metric on G (see Problems 3-10 and 3-12).

(a) For any $X, Y, Z \in \mathfrak{g}$, show that

$$\langle [X,Y], Z\rangle = -\langle Y, [X,Z]\rangle.$$

[Hint: Let $\gamma(t) = \exp(tX)$ (here exp denotes the Lie group exponential map, not the Riemannian one), and compute the t-derivative of $\langle \operatorname{Ad}_{\gamma(t)} Y, \operatorname{Ad}_{\gamma(t)} Z\rangle$ at $t = 0$, using the facts that $Ad_{\gamma(t)} = (R_{\gamma(-t)})_*(L_{\gamma(t)})_*$ and that $R_{\gamma(-t)}$ is the flow of $-X$.]

(b) Show that

$$\nabla_X Y = \frac{1}{2}[X, Y]$$

whenever X and Y are left-invariant vector fields on G.

(c) Show that the geodesics of g starting at the identity are exactly the one-parameter subgroups, so the Lie group exponential map coincides with the Riemannian exponential map at the identity.

6

Geodesics and Distance

In this chapter, we study in detail the relationships among geodesics, lengths, and distances on a Riemannian manifold. A primary goal is to show that all length-minimizing curves are geodesics, and that all geodesics are length minimizing, at least locally. A key ingredient in the proofs is the symmetry of the Riemannian connection. Later in the chapter, we study the property of geodesic completeness, which means that all maximal geodesics are defined for all time, and prove the Hopf–Rinow theorem, which states that a Riemannian manifold is geodesically complete if and only if it is complete as a metric space.

Throughout this chapter, M is a smooth n-manifold endowed with a fixed Riemannian metric g. All covariant derivatives and geodesics are understood to be with respect to the Riemannian connection of g.

Most of the results of this chapter do *not* apply to pseudo-Riemannian metrics, at least not without substantial modification. For a treatment of lengths of curves in the pseudo-Riemannian setting, see [O'N83].

Lengths and Distances on Riemannian Manifolds

We are now in a position to introduce two of the most fundamental concepts from classical geometry into the Riemannian setting: lengths of curves and distances between points. We begin with lengths.

Lengths of Curves

If $\gamma\colon [a,b] \to M$ is a curve segment, we define the *length* of γ to be

$$L(\gamma) := \int_a^b |\dot\gamma(t)|\,dt.$$

Sometimes, for the sake of clarity, we emphasize the dependence on the metric by using the notation L_g instead of L.

The key feature of the length of a curve is that it is independent of parametrization. To make this notion precise, we define a *reparametrization* of γ to be a curve segment of the form $\tilde\gamma = \gamma \circ \varphi$, where $\varphi\colon [c,d] \to [a,b]$ is a smooth map with smooth inverse. We say it is a *forward reparametrization* if φ is orientation preserving, and a *backward reparametrization* if not.

Lemma 6.1. *For any curve segment $\gamma\colon [a,b] \to M$, and any reparametrization $\tilde\gamma$ of γ, $L(\gamma) = L(\tilde\gamma)$.*

Exercise 6.1. Prove Lemma 6.1.

For measuring distances between points, it is useful to modify slightly the class of curves we consider. A *regular curve* is a smooth curve $\gamma\colon I \to M$ such that $\dot\gamma(t) \neq 0$ for $t \in I$. Intuitively, this prevents the curve from having "cusps" or "kinks." More formally, because the tangent vector $\dot\gamma(t)$ is the push-forward $\gamma_*(d/dt)$, a regular curve is an immersion of the interval I into M. (If I has one or two endpoints, it has to be considered as a manifold with boundary.) Note that geodesics are automatically regular, since they have constant speed.

A continuous map $\gamma\colon [a,b] \to M$ is called a *piecewise regular curve segment* if there exists a finite subdivision $a = a_0 < a_1 < \cdots < a_k = b$ such that $\gamma|_{[a_{i-1},a_i]}$ is a regular curve for $i = 1,\dots,k$. All distances on a Riemannian manifold will be measured along such curve segments. For brevity, we refer to a piecewise regular curve segment as an *admissible curve.* It's also convenient to allow a trivial constant curve $\gamma\colon \{a\} \to M$, $\gamma(a) = p$, to be considered an admissible curve.

The definition implies that an admissible curve must have well-defined, nonzero, one-sided velocity vectors when approaching a_i from either side, but the two limiting velocity vectors need not be equal. We denote these one-sided velocities by

$$\dot\gamma(a_i^-) := \lim_{t \nearrow a_i} \dot\gamma(t);$$

$$\dot\gamma(a_i^+) := \lim_{t \searrow a_i} \dot\gamma(t).$$

Let $\gamma\colon [a,b] \to M$ be an admissible curve, and $a = a_0 < a_1 < \cdots < a_k = b$ a subdivision as above. The length of γ is defined simply as the sum of the lengths of the smooth subsegments $\gamma|_{[a_{i-1},a_i]}$. We can broaden

the definition of reparametrization by defining a reparametrization of an admissible curve $\gamma\colon [a,b] \to M$ to be an admissible curve of the form $\tilde{\gamma} = \gamma \circ \varphi$, where $\varphi\colon [c,d] \to [a,b]$ is a homeomorphism whose restriction to each subinterval $[c_{i-1}, c_i]$ is smooth with smooth inverse, for some finite subdivision $c = c_0 < c_1 < \cdots < c_k = d$ of $[c,d]$. Then a straightforward generalization of Lemma 6.1 shows that the length of an admissible curve is also independent of parametrization.

The *arc length function* of an admissible curve $\gamma\colon [a,b] \to M$ is the function $s\colon [a,b] \to \mathbf{R}$ defined by

$$s(t) := L(\gamma|_{[a,t]}) = \int_a^t |\dot{\gamma}(u)|\, du.$$

It is an immediate consequence of the fundamental theorem of calculus that s is smooth wherever γ is, and $\dot{s}(t)$ is equal to the speed $|\dot{\gamma}(t)|$ of γ.

Among all the possible parametrizations of a given curve, the unit speed parametrizations are particularly useful. It is an important fact that every admissible curve has such a parametrization, as the next exercise shows.

Exercise 6.2. Let $\gamma\colon [a,b] \to M$ be an admissible curve, and set $l = L(\gamma)$.

(a) Show that there exists a unique forward reparametrization $\tilde{\gamma}\colon [0,l] \to M$ of γ such that $\tilde{\gamma}$ is a unit speed curve.

(b) If $\tilde{\gamma}$ is any unit speed curve whose parameter interval is of the form $[0,l]$, show that the arc length function of γ is $s(t) = t$. For this reason, such a curve is said to be *parametrized by arc length.*

If $\gamma\colon [a,b] \to M$ is any admissible curve, and $f \in C^\infty[a,b]$, we define the *integral of f with respect to arc length,* denoted $\int_\gamma f\, ds$, by

$$\int_\gamma f\, ds := \int_a^b f(t)\, |\dot{\gamma}(t)|\, dt.$$

Exercise 6.3. Let $\gamma\colon [a,b] \to M$ be an admissible curve, and $f \in C^\infty[a,b]$.

(a) Show that $\int_\gamma f\, ds$ is independent of parametrization.

(b) If γ is injective and smooth, show that $C := \gamma[a,b]$ is an embedded submanifold with boundary in M, and

$$\int_\gamma f\, ds = \int_C (f \circ \gamma^{-1})\, d\widetilde{V},$$

where $d\widetilde{V}$ is the Riemannian volume element on C associated with the induced metric and the orientation determined by γ.

A continuous map $V\colon [a,b] \to TM$ such that $V_t \in T_{\gamma(t)}M$ for all t is called a *piecewise smooth vector field along* γ if there is a (possibly finer) finite subdivision $a = \tilde{a}_0 < \tilde{a}_1 < \cdots < \tilde{a}_m = b$ such that V is smooth

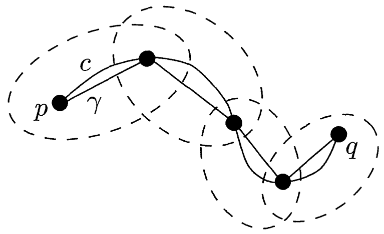


FIGURE 6.1. Any two points can be connected by an admissible curve.

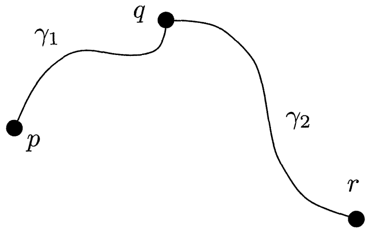


FIGURE 6.2. The triangle inequality.

on each subinterval $[\tilde{a}_{i-1}, \tilde{a}_i]$. Given any vector $V_a \in T_{\gamma(a)}M$, it is easy to check that V_a has a unique piecewise smooth parallel translate along all of γ; simply parallel translate V_a along the first smooth segment to $\gamma(a_1)$, then parallel translate V_{a_1} along the second smooth segment, and so on. The parallel translate is smooth wherever γ is.

The Riemannian Distance Function

Suppose M is a connected Riemannian manifold. For any pair of points $p, q \in M$, we define the *Riemannian distance* $d(p,q)$ to be the infimum of the lengths of all admissible curves from p to q. To check that this is well defined, we need to verify that any two points can be connected by an admissible curve. Since a connected manifold is path-connected, they can be connected by a continuous path $c\colon [a,b] \to M$. By compactness, there is a finite subdivision of $[a,b]$ such that $c[a_{i-1}, a_i]$ is contained in a single chart for each i. Then we may replace each such segment by a smooth path in coordinates yielding an admissible curve γ between the same points (Figure 6.1). Therefore $d(p,q)$ is finite for each $p, q \in M$.

Lemma 6.2. *With the distance function d defined above, any connected Riemannian manifold is a metric space whose induced topology is the same as the given manifold topology.*

Proof. It is obvious from the definition that $d(p,q) = d(q,p) \geq 0$ and $d(p,p) = 0$. The triangle inequality follows from the fact that an admissible curve from p to q can be combined with one from q to r (possibly changing the starting time of the parametrization of the second) to yield one from p to r whose length is the sum of the lengths of the two given curves (Figure 6.2). (This is one reason for defining distance using *piecewise* regular curves instead of just regular ones.)

It remains to show that $d(p,q) > 0$ when $p \neq q$, and that the metric topology is the same as the manifold topology. To do so, we need to compare the Riemannian distance to the Euclidean distance in local coordinates. Let

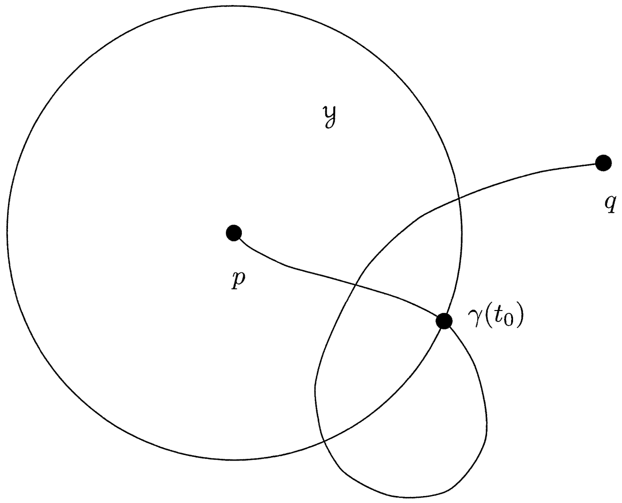


FIGURE 6.3. $d(p,q) > 0$.

$p \in M$, and let (x^i) be normal coordinates centered at p. Arguing as in the proof of the uniformly normal neighborhood lemma (Lemma 5.12), there exists a closed geodesic ball $\overline{\mathcal{Y}}$ of radius ε around p and positive constants c and C such that $c|V|_{\bar g} \le |V|_g \le C|V|_{\bar g}$ whenever $V \in T_xM$ and $x \in \overline{\mathcal{Y}}$. It follows immediately from the definition of length that for any admissible curve γ whose image is contained in $\overline{\mathcal{Y}}$,

$$cL_{\bar g}(\gamma) \le L_g(\gamma) \le CL_{\bar g}(\gamma). \tag{6.1}$$

Now if $q \ne p$, we may shrink ε so that $q \notin \overline{\mathcal{Y}}$. Then any admissible curve $\gamma\colon [a,b] \to M$ from p to q must pass through the geodesic sphere $\partial\mathcal{Y}$ (since the complement of the sphere is disconnected, and p,q lie in different components). If we let t_0 denote the first such time (Figure 6.3), it follows that

$$d(p,q) \ge L_g(\gamma) \ge L_g(\gamma|_{[a,t_0]}) \ge cL_{\bar g}(\gamma|_{[a,t_0]}) \ge c\, d_{\bar g}(p,\gamma(t_0)) = c\varepsilon > 0.$$

Thus d is a metric.

Finally, to compare the two topologies, just note that we can construct a basis for the manifold topology from small Euclidean balls in open sets of the form $\mathcal{Y}$ as above, and the metric topology is generated by small metric balls. The discussion above shows that in any such set $\mathcal{Y}$, the Euclidean distance and the Riemannian distance are equivalent, so the basis open sets in either topology are open in both. This shows that the two topologies are the same. □

Geodesics and Minimizing Curves

An admissible curve γ in a Riemannian manifold is said to be *minimizing* if $L(\gamma) \leq L(\tilde{\gamma})$ for any other admissible curve $\tilde{\gamma}$ with the same endpoints. It follows immediately from the definition of distance that γ is minimizing if and only if $L(\gamma)$ is equal to the distance between its endpoints.

To show that all minimizing curves are geodesics, we will think of the length function L as a functional on the set of admissible curves in M. (Functions whose domains are themselves sets of functions are usually called "functionals.") From this point of view, the search for minimizing curves can be thought of as searching for minima of this functional.

From calculus, we might expect that a necessary condition for a curve γ to be minimizing would be that the "derivative" of L vanish at γ, in some sense. This brings us to the brink of the subject known as the *calculus of variations:* the use of calculus to identify and analyze extrema of functionals defined on spaces of functions or maps. In its fully developed state, the calculus of variations allows one to apply all the usual tools of multivariable calculus in the infinite-dimensional setting of function spaces, such as directional derivatives, gradients, critical points, local extrema, saddle points, and Hessians. For our purposes, however, we do not need to formalize the theory of calculus in the infinite-dimensional setting. It suffices to note that if γ is a minimizing curve, and Γ_s is a family of admissible curves with the same endpoints such that $L(\Gamma_s)$ is a differentiable function of s and $\Gamma_0 = \gamma$, then by elementary calculus $L(\Gamma_s)$ must have vanishing s-derivative at $s = 0$ because it attains a minimum there.

Admissible Families

To make this rigorous, we introduce some more definitions. An *admissible family* of curves is a continuous map $\Gamma\colon (-\varepsilon, \varepsilon) \times [a, b] \to M$ that is smooth on each rectangle of the form $(-\varepsilon, \varepsilon) \times [a_{i-1}, a_i]$ for some finite subdivision $a = a_0 < \cdots < a_k = b$, and such that $\Gamma_s(t) := \Gamma(s, t)$ is an admissible curve for each $s \in (-\varepsilon, \varepsilon)$ (Figure 6.4). If Γ is an admissible family, a *vector field along* Γ is a continuous map $V\colon (-\varepsilon, \varepsilon) \times [a, b] \to TM$ such that $V(s, t) \in T_{\Gamma(s,t)}M$ for each (s, t), and such that $V|_{(-\varepsilon,\varepsilon)\times[\tilde{a}_{i-1},\tilde{a}_i]}$ is smooth for some (possibly finer) subdivision $a = \tilde{a}_0 < \cdots < \tilde{a}_m = b$.

Any admissible family Γ defines two collections of curves: the *main curves* $\Gamma_s(t) = \Gamma(s, t)$ defined on $[a, b]$ by setting $s =$ constant, and the *transverse curves* $\Gamma^{(t)}(s) = \Gamma(s, t)$ defined on $(-\varepsilon, \varepsilon)$ by setting $t =$ constant. The transverse curves are smooth on $(-\varepsilon, \varepsilon)$ for each t, while the main curves are in general only piecewise regular. Wherever Γ is smooth, the tangent vectors to these two families of curves are examples of vector fields along

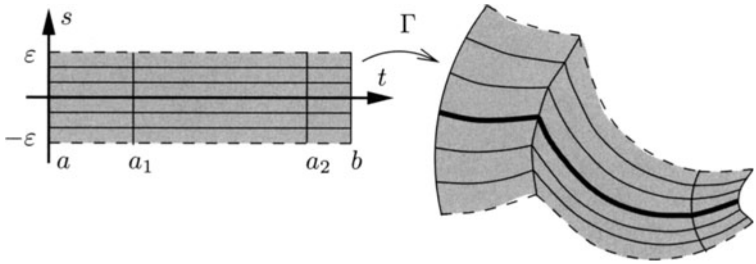


FIGURE 6.4. An admissible family.

Γ; we denote them by

$$\partial_t\Gamma(s,t) := \frac{d}{dt}\Gamma_s(t); \qquad \partial_s\Gamma(s,t) := \frac{d}{ds}\Gamma^{(t)}(s).$$

In fact, $\partial_s\Gamma$ is always continuous on the whole rectangle $(-\varepsilon,\varepsilon)\times[a,b]$: on one hand, its value along the line segment $(-\varepsilon,\varepsilon)\times\{a_i\}$ depends only on the values of Γ on that segment, since the derivative is taken only with respect to the s variable; on the other hand, it is continuous (in fact smooth) on each subrectangle $(-\varepsilon,\varepsilon)\times[a_{i-1},a_i]$ and $(-\varepsilon,\varepsilon)\times[a_i,a_{i+1}]$, so the right-handed and left-handed limits at $t = a_i$ must be equal. Therefore $\partial_s\Gamma$ is always a vector field along Γ. (However, $\partial_t\Gamma$ is not usually continuous at $t = a_i$.)

If V is a vector field along Γ, we can compute the covariant derivative of V either along the main curves or along the transverse curves, at least where the former are smooth; the resulting vector fields along Γ are denoted D_tV and D_sV respectively.

As mentioned earlier, a key ingredient in the proof that minimizing curves are geodesics is the symmetry of the Riemannian connection. It enters into our proofs in the form of the following lemma. (Although we state and use this lemma only for the Riemannian connection, the proof shows that it is actually true for any symmetric connection.)

Lemma 6.3. (Symmetry Lemma) *Let* $\Gamma\colon(-\varepsilon,\varepsilon)\times[a,b]\to M$ *be an admissible family of curves in a Riemannian (or pseudo-Riemannian) manifold. On any rectangle* $(-\varepsilon,\varepsilon)\times[a_{i-1},a_i]$ *where* Γ *is smooth,*

$$D_s\partial_t\Gamma = D_t\partial_s\Gamma.$$

Proof. This is a local question, so we may compute in coordinates (x^i) around any point $\Gamma(s_0,t_0)$. Writing the components of Γ as $\Gamma(s,t) = (x^1(s,t),\ldots,x^n(s,t))$, we have

$$\partial_t\Gamma = \frac{\partial x^k}{\partial t}\partial_k; \qquad \partial_s\Gamma = \frac{\partial x^k}{\partial s}\partial_k.$$

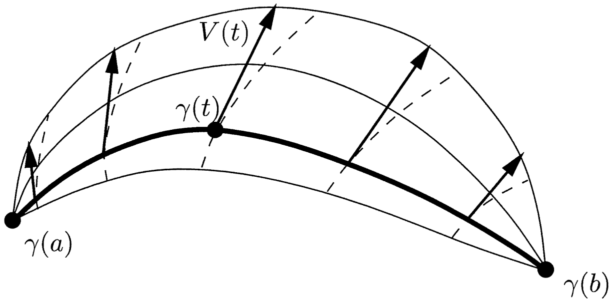


FIGURE 6.5. Every vector field along γ is the variation field of a variation of γ.

Then, using the coordinate formula (4.10) for covariant derivatives along curves,

$$D_s\partial_t\Gamma = \left(\frac{\partial^2 x^k}{\partial s\partial t} + \frac{\partial x^i}{\partial t}\frac{\partial x^j}{\partial s}\Gamma^k_{ji}\right)\partial_k;$$

$$D_t\partial_s\Gamma = \left(\frac{\partial^2 x^k}{\partial t\partial s} + \frac{\partial x^i}{\partial s}\frac{\partial x^j}{\partial t}\Gamma^k_{ji}\right)\partial_k.$$

Reversing the roles of i and j in the second line above, and using the symmetry condition $\Gamma^k_{ji} = \Gamma^k_{ij}$, we see immediately that these two expressions are equal. □

If $\gamma\colon [a,b] \to M$ is an admissible curve, a *variation* of γ is an admissible family Γ such that $\Gamma_0(t) = \gamma(t)$ for all $t \in [a,b]$. It is called a *proper variation* or *fixed-endpoint variation* if in addition $\Gamma_s(a) = \gamma(a)$ and $\Gamma_s(b) = \gamma(b)$ for all s. If Γ is a variation of γ, the *variation field* of Γ is the vector field $V(t) = \partial_s\Gamma(0,t)$ along γ. A vector field V along γ is *proper* if $V(a) = V(b) = 0$. It is clear that the variation field of a proper variation is itself proper.

Lemma 6.4. *If γ is an admissible curve and V is a vector field along γ, then V is the variation field of some variation of γ. If V is proper, the variation can be taken to be proper as well.*

Proof. Set $\Gamma(s,t) = \exp(sV(t))$ (Figure 6.5). By compactness of $[a,b]$, there is some positive ε such that Γ is defined on $(-\varepsilon,\varepsilon) \times [a,b]$. Clearly Γ is smooth on $(-\varepsilon,\varepsilon) \times [a_{i-1},a_i]$ for each subinterval $[a_{i-1},a_i]$ on which V is smooth, and is continuous on its whole domain. By the properties of the exponential map, the variation field of Γ is V. Moreover, if $V(a) = V(b) = 0$, it is immediate that $\Gamma(s,a) \equiv \gamma(a)$ and $\Gamma(s,b) \equiv \gamma(b)$, so Γ is proper. □

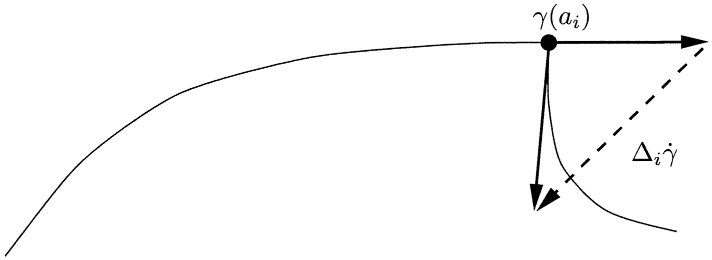


FIGURE 6.6. $\Delta_i\dot\gamma$ is the "jump" in $\dot\gamma$ at a_i.

Minimizing Curves Are Geodesics

We can now compute an expression for the derivative of the length functional along a proper variation. Traditionally, the derivative of a functional on a space of maps is called its *first variation.*

Proposition 6.5. (First Variation Formula) *Let $\gamma\colon [a,b] \to M$ be any unit speed admissible curve, Γ a proper variation of γ, and V its variation field. Then*

$$\left.\frac{d}{ds}\right|_{s=0} L(\Gamma_s) = -\int_a^b \langle V, D_t\dot\gamma\rangle\, dt - \sum_{i=1}^{k-1}\langle V(a_i), \Delta_i\dot\gamma\rangle, \tag{6.2}$$

where $\Delta_i\dot\gamma = \dot\gamma(a_i^+) - \dot\gamma(a_i^-)$ is the "jump" in the tangent vector field $\dot\gamma$ at a_i (Figure 6.6).

Proof. For brevity, denote

$$T(s,t) = \partial_t\Gamma(s,t), \qquad S(s,t) = \partial_s\Gamma(s,t).$$

On any subinterval $[a_{i-1}, a_i]$ where Γ is smooth, since the integrand in $L(\Gamma_s)$ is smooth and the domain of integration is compact, we can differentiate under the integral sign to obtain

$$\begin{aligned}\frac{d}{ds}L(\Gamma_s|_{[a_{i-1},a_i]}) &= \int_{a_{i-1}}^{a_i} \frac{\partial}{\partial s}\langle T,T\rangle^{1/2}\, dt \\ &= \int_{a_{i-1}}^{a_i} \frac{1}{2}\langle T,T\rangle^{-1/2}\, 2\,\langle D_sT, T\rangle\; dt \\ &= \int_{a_{i-1}}^{a_i} \frac{1}{|T|}\,\langle D_tS, T\rangle\; dt,\end{aligned} \tag{6.3}$$

where we have used the symmetry lemma in the last line. Setting $s = 0$ and noting that $S(0,t) = V(t)$ and $T(0,t) = \dot\gamma(t)$ (which has length 1),

$$\begin{aligned}\left.\frac{d}{ds}\right|_{s=0} L(\Gamma_s|_{[a_{i-1},a_i]}) &= \int_{a_{i-1}}^{a_i} \langle D_t V, \dot\gamma\rangle \, dt \\ &= \int_{a_{i-1}}^{a_i} \left(\frac{d}{dt}\langle V, \dot\gamma\rangle - \langle V, D_t\dot\gamma\rangle\right) dt \\ &= \langle V(a_i), \dot\gamma(a_i^-)\rangle - \langle V(a_{i-1}), \dot\gamma(a_{i-1}^+)\rangle \\ &\qquad - \int_{a_{i-1}}^{a_i} \langle V, D_t\dot\gamma\rangle \, dt.\end{aligned}$$

Finally, summing over i and noting $V(a_0) = V(a_k) = 0$ because Γ is a proper variation, we obtain (6.2). □

Because any admissible curve has a unit speed parametrization and length is independent of parametrization, the requirement in the above proposition that γ be unit speed is not a real restriction, but rather just a computational convenience.

Exercise 6.4. Let γ be a smooth, unit speed curve.

(a) Show that $D_t\dot\gamma(t)$ is orthogonal to $\dot\gamma(t)$ for all t.

(b) If Γ is a proper variation of γ such that for all s, Γ_s is a reparametrization of γ, show that the first variation of $L(\Gamma_s)$ vanishes.

Theorem 6.6. *Every minimizing curve is a geodesic when it is given a unit speed parametrization.*

Proof. Suppose $\gamma\colon [a,b] \to M$ is minimizing and unit speed, and let $a = a_0 < \cdots < a_k = b$ be a subdivision such that γ is smooth on $[a_{i-1}, a_i]$. If Γ is any proper variation of γ, we conclude from elementary calculus that $dL(\Gamma_s)/ds = 0$ when $s = 0$. Since every proper vector field along γ is the variation field of some proper variation, the right-hand side of (6.2) must vanish for every such V.

The first step is to show that $D_t\dot\gamma = 0$ on each subinterval $[a_{i-1}, a_i]$, so γ is a "broken geodesic." Choose one such interval, and let $\varphi \in C^\infty(\mathbf{R})$ be a bump function such that $\varphi > 0$ on (a_{i-1}, a_i) and $\varphi = 0$ elsewhere. Then (6.2) with $V = \varphi D_t\dot\gamma$ becomes

$$0 = -\int_{a_{i-1}}^{a_i} \varphi \left|D_t\dot\gamma\right|^2 dt.$$

Since the integrand is nonnegative, this shows that $D_t\dot\gamma = 0$ on each such subinterval.

Next we need to show that $\Delta_i\dot\gamma = 0$, which is to say that γ has no corners. For any i between 0 and k, it is easy to use a bump function in a coordinate

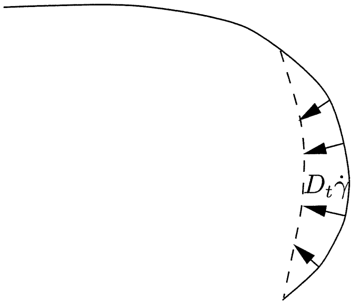


FIGURE 6.7. Deforming γ in the direction of its acceleration vector.

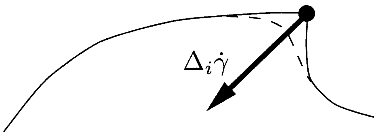


FIGURE 6.8. Rounding the corner.

chart to construct a vector field V along γ such that $V(a_i) = \Delta_i\dot\gamma$ and $V(a_j) = 0$ for $j \neq i$. Then (6.2) reduces to $-|\Delta_i\dot\gamma|^2 = 0$.

Finally, since the two one-sided velocity vectors of γ match up at each a_i, it follows from uniqueness of geodesics that $\gamma|_{[a_i,a_{i+1}]}$ is the continuation of the geodesic $\gamma|_{[a_{i-1},a_i]}$, and therefore γ is smooth. □

The preceding proof has an enlightening geometric interpretation. Assuming $D_t\dot\gamma \neq 0$, the first variation with $V = \varphi D_t\dot\gamma$ is negative, which shows that deforming γ in the direction of its acceleration vector decreases its length (Figure 6.7). Similarly, the length of a broken geodesic γ is decreased by deforming it in the direction of a vector field V such that $V(a_i) = \Delta_i\dot\gamma$ (Figure 6.8). Geometrically, this corresponds to "rounding the corner."

The first variation formula actually tells us a bit more than is claimed in Theorem 6.6. In proving that γ is a geodesic, we didn't use the full strength of the assumption that it is a minimizing curve—we used only the fact that it is a *critical point* of L, which means that for any proper variation Γ_s of γ, the derivative of $L(\Gamma_s)$ with respect to s is zero at $s = 0$. Therefore we can strengthen Theorem 6.6 in the following way.

Corollary 6.7. *A unit speed admissible curve γ is a critical point for L if and only if it is a geodesic.*

Proof. If γ is a critical point, the proof of Theorem 6.6 goes through without modification to show that γ is a geodesic. Conversely, if γ is a geodesic, then the first term in the second variation formula vanishes by the geodesic equation, and the second term vanishes because $\dot\gamma$ has no jumps. □

The geodesic equation $D_t\dot\gamma = 0$ thus characterizes the critical points of the length functional. In general, the equation that characterizes critical points of a functional on a space of maps is called the *variational equation* or the *Euler–Lagrange equation* of the functional. Many interesting equations

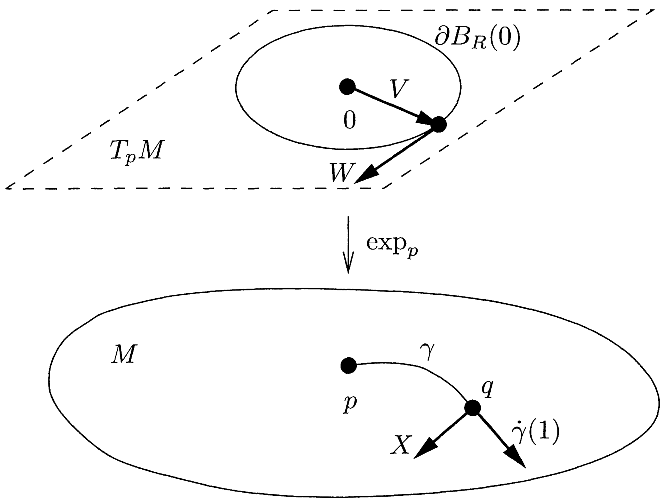


FIGURE 6.9. Proof of the Gauss lemma.

in differential geometry arise as variational equations. We touch briefly on three others in this book: the Einstein equation (Chapter 7), the Yamabe equation (Chapter 7), and the minimal surface equation (Chapter 8).

Geodesics Are Locally Minimizing

Next we turn to the converse of Theorem 6.6, and show that geodesics are locally minimizing. The proof is based on the following deceptively simple geometric fact.

Theorem 6.8. (The Gauss Lemma) *Let $\mathcal{U}$ be a geodesic ball centered at $p \in M$. The unit radial vector field $\partial/\partial r$ is g-orthogonal to the geodesic spheres in $\mathcal{U}$.*

Proof. Let $q \in \mathcal{U}$ and let $X \in T_qM$ be a vector tangent to the geodesic sphere through q. Because $\exp_p$ is a diffeomorphism onto $\mathcal{U}$, there is a vector $V \in T_pM$ such that $q = \exp_p V$, and there is a vector $W \in T_V(T_pM) = T_pM$ such that $X = (\exp_p)_* W$ (Figure 6.9). Then $V \in \partial B_R(0)$ and $W \in T_V \partial B_R(0)$, where $R = d(p,q)$. The radial geodesic from p to q is $\gamma(t) = \exp_p(tV)$, with tangent vector $\dot\gamma(t) = R\partial/\partial r$. Thus we need to show that $X \perp \dot\gamma(1)$ with respect to g.

Choose a curve $\sigma\colon (-\varepsilon, \varepsilon) \to T_pM$ lying in $\partial B_R(0)$ such that $\sigma(0) = V$ and $\dot\sigma(0) = W$, and consider the variation Γ of γ (Figure 6.10) given by

$$\Gamma(s,t) = \exp_p(t\sigma(s)).$$

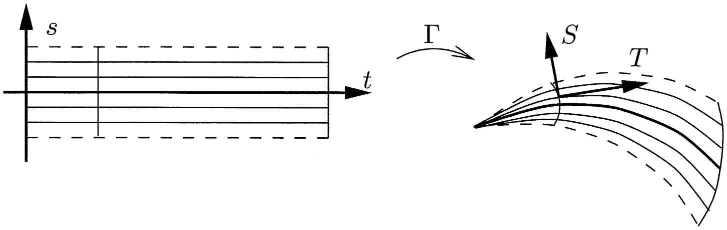


FIGURE 6.10. The variation Γ.

For each $s \in (-\varepsilon, \varepsilon)$, $\sigma(s)$ is a vector of length R, so Γ_s is a geodesic with constant speed R. As before, let $S = \partial_s \Gamma$ and $T = \partial_t \Gamma$. It follows from the definitions that

$$\begin{aligned} S(0,0) &= \left.\frac{d}{ds}\right|_{s=0} \exp_p(0) = 0; \\ T(0,0) &= \left.\frac{d}{dt}\right|_{t=0} \exp_p(tV) = V; \\ S(0,1) &= \left.\frac{d}{ds}\right|_{s=0} \exp_p(\sigma(s)) = (\exp_p)_* \dot\sigma(0) = X; \\ T(0,1) &= \left.\frac{d}{dt}\right|_{t=1} \exp_p(tV) = \dot\gamma(1). \end{aligned}$$

Therefore $\langle S, T \rangle$ is zero when $(s,t) = (0,0)$ and equal to $\langle X, \dot\gamma(1) \rangle$ when $(s,t) = (0,1)$, so to prove the theorem it suffices to show $\langle S, T \rangle$ is independent of t.

We compute

$$\begin{aligned} \frac{\partial}{\partial t} \langle S, T \rangle &= \langle D_t S, T \rangle + \langle S, D_t T \rangle \\ &= \langle D_s T, T \rangle + 0 \\ &= \frac{1}{2} \frac{\partial}{\partial s} |T|^2 = 0, \end{aligned}$$

where we have used (1) the symmetry lemma $D_t S = D_s T$, (2) the fact that $D_t T \equiv 0$ since each Γ_s is a geodesic, and (3) the fact that $|T| = |\dot\Gamma_s| \equiv R$ for all (s,t). This proves the theorem. □

We will use the Gauss lemma primarily in the form of the next corollary.

Corollary 6.9. *Let (x^i) be normal coordinates on a geodesic ball $\mathcal{U}$ centered at $p \in M$, and let r be the radial distance function as defined in* (5.9). *Then* $\operatorname{grad} r = \partial/\partial r$ *on* $\mathcal{U} - \{p\}$.

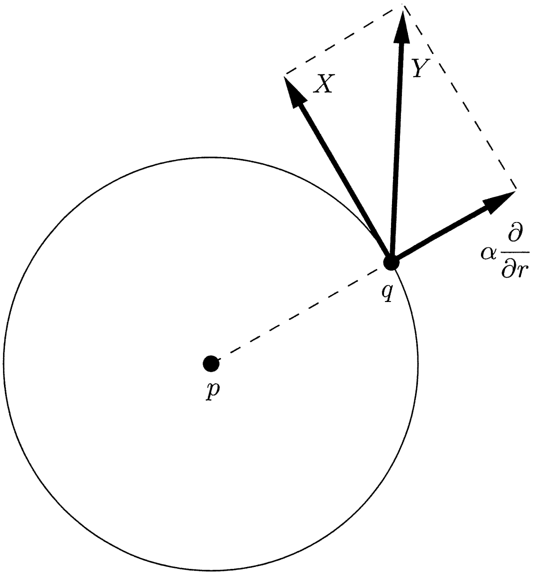


FIGURE 6.11. Decomposition of Y into tangential and normal components.

Proof. For any $q \in \mathcal{U} - \{p\}$ and $Y \in T_qM$, we need to show that

$$dr(Y) = \left\langle \frac{\partial}{\partial r}, Y \right\rangle. \tag{6.4}$$

The geodesic sphere $\exp_p(\partial B_R(0))$ through q is characterized in normal coordinates by the equation $r = R$. Since $\partial/\partial r$ is transverse to this sphere, we can decompose Y as $\alpha\, \partial/\partial r + X$ for some constant α and some vector X tangent to the sphere (Figure 6.11). Observe that $dr(\partial/\partial r) = 1$ by direct computation in coordinates, and $dr(X) = 0$ since X is tangent to a level set of r. (This has nothing to do with the metric!) Therefore the left-hand side of (6.4) is

$$dr\left(\alpha \frac{\partial}{\partial r} + X\right) = \alpha\, dr\left(\frac{\partial}{\partial r}\right) + dr(X) = \alpha.$$

On the other hand, by Proposition 5.11(e), $\partial/\partial r$ is a unit vector. Therefore, the right-hand side of (6.4) is

$$\left\langle \frac{\partial}{\partial r}, \alpha \frac{\partial}{\partial r} + X \right\rangle = \alpha \left| \frac{\partial}{\partial r} \right|^2 + \left\langle \frac{\partial}{\partial r}, X \right\rangle = \alpha,$$

where we have used the Gauss lemma to conclude that X is orthogonal to $\partial/\partial r$. □

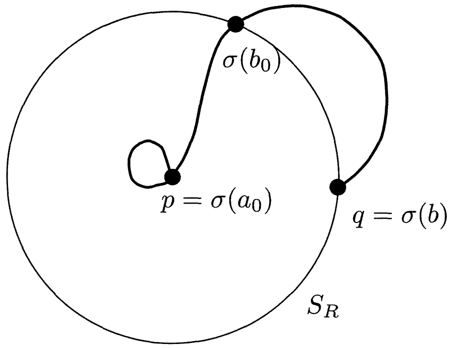


FIGURE 6.12. Radial geodesics are minimizing.

Proposition 6.10. *Suppose $p \in M$ and q is contained in a geodesic ball around p. Then (up to reparametrization) the radial geodesic from p to q is the unique minimizing curve from p to q in M.*

Proof. Choose $\varepsilon > 0$ such that $\exp_p(B_\varepsilon(0))$ is a geodesic ball containing q. Let $\gamma\colon [0, R] \to M$ be the radial geodesic from p to q parametrized by arc length, and write $\gamma(t) = \exp_p(tV)$ for some unit vector $V \in T_pM$. Then $L(\gamma) = R$ since γ has unit speed, so we need to show that any other admissible curve from p to q has length strictly greater than R. Let $S_R = \exp_p(\partial B_R(0))$ denote the geodesic sphere of radius R.

Let $\sigma\colon [0, b] \to M$ be such a curve, which we may assume to be parametrized by arc length as well. We begin by showing $L(\sigma) \geq L(\gamma)$.

Let $a_0 \in [a, b]$ denote the last time that $\sigma(t) = p$ and $b_0 \in [a, b]$ the first time after a_0 that $\sigma(t) \in S_R$ (Figure 6.12). For any $t \in (a_0, b_0]$, we can decompose $\dot\sigma(t)$ as

$$\dot\sigma(t) = \alpha(t)\frac{\partial}{\partial r} + X(t),$$

where $X(t)$ is tangent to the geodesic sphere through $\sigma(t)$. By the Gauss lemma, this is an orthogonal decomposition, so $|\dot\sigma(t)|^2 = \alpha(t)^2 + |X(t)|^2 \geq$

$\alpha(t)^2$. Moreover, by Corollary 6.9, $\alpha(t) = \langle \partial/\partial r, \dot\sigma(t)\rangle = dr(\dot\sigma(t))$. Therefore

$$\begin{aligned}
L(\sigma) &\geq L(\sigma|_{[a_0,b_0]}) \\
&= \lim_{\delta\to 0} \int_{a_0+\delta}^{b_0} |\dot\sigma(t)|\, dt \\
&\geq \lim_{\delta\to 0} \int_{a_0+\delta}^{b_0} \alpha(t)\, dt \\
&= \lim_{\delta\to 0} \int_{a_0+\delta}^{b_0} dr(\dot\sigma(t))\, dt \\
&= \lim_{\delta\to 0} \int_{a_0+\delta}^{b_0} \frac{d}{dt} r(\sigma(t))\, dt \\
&= r(\sigma(b_0)) - r(\sigma(a_0)) \\
&= R = L(\gamma).
\end{aligned} \tag{6.5}$$

Thus γ is minimizing.

Now suppose $L(\sigma) = R$. Then both inequalities in (6.5) are equalities. Because we assume σ is a unit speed curve, the first equality implies that $a_0 = 0$ and $b_0 = b = R$, since otherwise the segments of σ before $t = a_0$ and after $t = b_0$ would contribute positive lengths. The second equality implies that $X(t) \equiv 0$ and $\alpha(t) > 0$, so $\dot\sigma(t)$ is a positive multiple of $\partial/\partial r$. For σ to have unit speed we must have $\dot\sigma(t) = \partial/\partial r$. Thus σ and γ are both integral curves of $\partial/\partial r$ passing through q at time $t = R$, so $\sigma = \gamma$. □

Corollary 6.11. *Within any geodesic ball around $p \in M$, the radial distance function $r(x)$ defined by* (5.9) *is equal to the Riemannian distance from p to x.*

Proof. The radial geodesic γ from p to x is minimizing by Proposition 6.10. Since its velocity is equal to $\partial/\partial r$, which is a unit vector in both the g norm and the Euclidean norm in normal coordinates, the g-length of γ is equal to its Euclidean length, which is $r(x)$. □

This corollary suggests a simplified notation for geodesic balls and spheres in M. If $\mathcal{U} = \exp_p(B_R(0))$ is a geodesic ball around p, Corollary 6.11 shows that $\mathcal{U}$ is equal to the metric ball of radius R around p. Similarly, a geodesic sphere of radius R is the set of points whose distance from p is exactly R. From now on, we will use the notations $B_R(p) = \exp_p(B_R(0))$, $\overline{B}_R(p) = \exp_p(\overline{B}_R(0))$, and $S_R(p) = \exp_p(\partial B_R(0))$ for open and closed geodesic balls and geodesic spheres, which are exactly those metric balls and spheres that lie within a normal neighborhood of p.

We say a curve $\gamma\colon I \to M$ is *locally minimizing* if any $t_0 \in I$ has a neighborhood $\mathcal{U} \subset I$ such that $\gamma|_{\mathcal{U}}$ is minimizing between each pair of its points. Note that a minimizing curve is automatically locally minimizing, because it is minimizing between any two of its points.

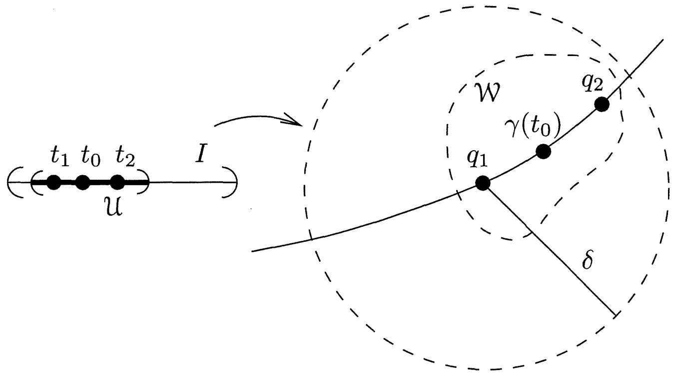


FIGURE 6.13. Geodesics are locally minimizing.

Theorem 6.12. *Every Riemannian geodesic is locally minimizing.*

Proof. Let $\gamma\colon I \to M$ be a geodesic, which we may assume to be defined on an open interval, and let $t_0 \in I$. Let $\mathcal{W}$ be a uniformly normal neighborhood of $\gamma(t_0)$, and let $\mathcal{U} \subset I$ be the connected component of $\gamma^{-1}(\mathcal{W})$ containing t_0. If $t_1, t_2 \in \mathcal{U}$ and $q_i = \gamma(t_i)$, the definition of uniformly normal neighborhood implies that q_2 is contained in a geodesic ball around q_1 (Figure 6.13). Therefore, by Proposition 6.10, the radial geodesic from q_1 to q_2 is the unique minimizing curve between them. However, the restriction of γ is a geodesic from q_1 to q_2 lying in the same geodesic ball, and thus γ must itself be this minimizing geodesic. □

It is interesting to note that the Gauss lemma and its corollary also yield another proof that minimizing curves are geodesics, without using the first variation formula. On the principle that knowing more than one proof of an important fact always deepens our understanding of it, we present this proof for good measure.

Another proof of Theorem 6.6. Suppose $\gamma\colon [a,b] \to M$ is any minimizing curve segment. Just as in the preceding proof, for any $t_0 \in [a,b]$ we can find a connected neighborhood $\mathcal{U}$ of t_0 such that $\gamma(\mathcal{U})$ is contained in a uniformly normal neighborhood $\mathcal{W}$. Then for any $t_1, t_2 \in \mathcal{U}$, the same argument as above shows that the unique minimizing curve from $\gamma(t_1)$ to $\gamma(t_2)$ is the radial geodesic joining them. Since the restriction of γ is such a minimizing curve, it must coincide with this radial geodesic. Therefore γ solves the geodesic equation in a neighborhood of t_0. Since t_0 was arbitrary, γ is a geodesic. □

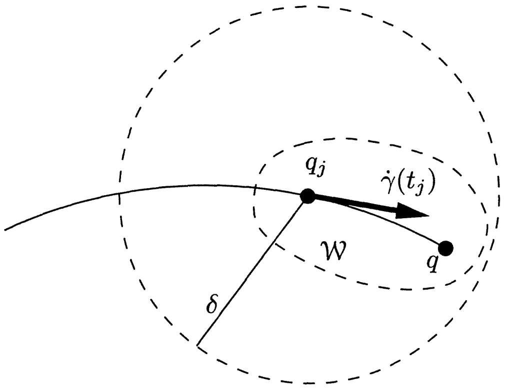


FIGURE 6.14. γ extends past q.

Completeness

A Riemannian manifold is said to be *geodesically complete* if every maximal geodesic is defined for all $t \in \mathbf{R}$. It is easy to construct examples of manifolds that are not geodesically complete; for example, in any proper open subset of $\mathbf{R}^n$ with its Euclidean metric, there are geodesics that reach the boundary in finite time. Similarly, on $\mathbf{R}^n$ with the metric $(\sigma^{-1})^*\mathring{g}$ obtained from the sphere by stereographic projection, there are geodesics that escape to infinity in finite time. The following theorem provides a simple criterion for determining when a Riemannian manifold is geodesically complete.

Theorem 6.13. (Hopf–Rinow) *A connected Riemannian manifold is geodesically complete if and only if it is complete as a metric space.*

Proof. Suppose first that M is complete as a metric space but not geodesically complete. Then there is some unit speed geodesic $\gamma\colon [0,b) \to M$ that extends to no interval $[0, b+\varepsilon)$ for $\varepsilon > 0$. Let $\{t_i\}$ be any increasing sequence that approaches b, and set $q_i = \gamma(t_i)$. Since γ is parametrized by arc length, the length of $\gamma|_{[t_i,t_j]}$ is exactly $|t_j - t_i|$, so $d(q_i, q_j) \le |t_j - t_i|$ and $\{q_i\}$ is a Cauchy sequence in M. By completeness, $\{q_i\}$ converges to some point $q \in M$.

Let $\mathcal{W}$ be a uniformly normal neighborhood of q, and let $\delta > 0$ be chosen so that $\mathcal{W}$ is contained in a geodesic δ-ball around each of its points. For all large j, $q_j \in \mathcal{W}$ (Figure 6.14), and by taking j large enough, we may assume $t_j > b - \delta$. The fact that $B_\delta(q_j)$ is a geodesic ball means that every geodesic starting at q_j exists at least for time δ. In particular, this is true of the geodesic σ with $\sigma(0) = q_j$ and $\dot\sigma(0) = \dot\gamma(t_j)$. But by uniqueness of geodesics, this must be simply a reparametrization of γ, so $\tilde\gamma(t) = \sigma(t_j + t)$ is an extension of γ past b, which is a contradiction.

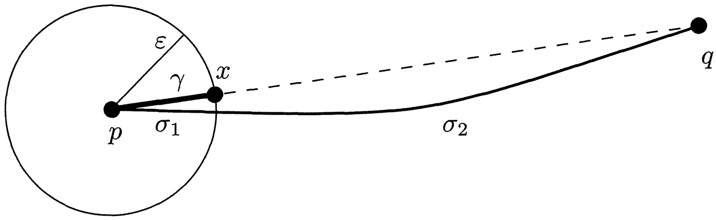


FIGURE 6.15. Proof that $\gamma|_{[0,\varepsilon]}$ aims at q.

To prove the converse, we will actually prove something stronger: If there is one point $p \in M$ such that $\exp_p$ is defined on the whole tangent space T_pM, then M is a complete metric space.

Suppose p is such a point. We show first that given any other point $q \in M$, there is a minimizing geodesic segment from p to q. If $\gamma\colon [0,b] \to M$ is a geodesic segment, we say that γ *aims at* q if γ is minimizing and

$$d(\gamma(0), q) = d(\gamma(0), \gamma(b)) + d(\gamma(b), q). \tag{6.6}$$

(This, of course, would be the case if γ were an initial segment of a minimizing geodesic from $\gamma(0)$ to q.) It will suffice to show that there is a geodesic segment γ that begins at p, aims at q, and has length equal to $d(p,q)$, for then (6.6) says that

$$d(p,q) = d(p,q) + d(\gamma(b), q),$$

which implies $\gamma(b) = q$. Since γ is assumed to be minimizing, it is the desired geodesic segment.

Choose $\varepsilon > 0$ such that $\overline{B}_\varepsilon(p)$ is a closed geodesic ball around p. If $q \in \overline{B}_\varepsilon(p)$, there is a minimizing geodesic from p to q by Proposition 6.10, and we have nothing more to prove. If $q \notin \overline{B}_\varepsilon(p)$, since the distance function on any metric space is continuous there is a point $x \in S_\varepsilon(p)$ where $d(x,q)$ attains its minimum on the compact set $S_\varepsilon(p)$. Let γ be the unit speed radial geodesic from p to x (Figure 6.15); by assumption, γ is defined for all time.

We begin by showing that $\gamma|_{[0,\varepsilon]}$ aims at q. Since it is minimizing by Proposition 6.10, we need only show that (6.6) holds with $b = \varepsilon$, or $d(p,q) = d(p,x) + d(x,q)$. By the triangle inequality, the only way for this to fail is if $d(p,q) < d(p,x) + d(x,q)$. Then there is a unit speed admissible curve σ from p to q whose length is strictly less than $d(p,x) + d(x,q)$. Let σ_1 denote the portion of σ inside $\overline{B}_\varepsilon(p)$, and σ_2 the rest (Figure 6.15). Then, since $L(\sigma_1) \geq \varepsilon$,

$$\begin{aligned} d(p,x) + d(x,q) &> L(\sigma) \\ &\geq \varepsilon + L(\sigma_2) \\ &= d(p,x) + L(\sigma_2). \end{aligned}$$

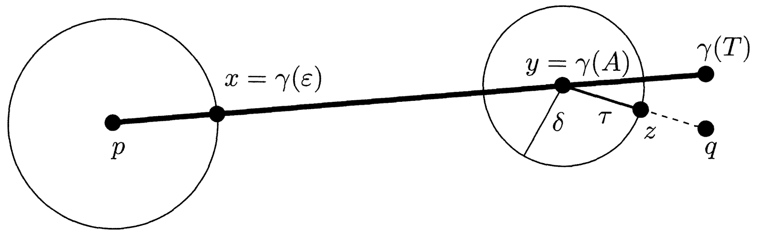


FIGURE 6.16. Proof that $A = T$.

But this means $L(\sigma_2) < d(x,q)$, which contradicts our choice of x.

Let $T = d(p,q)$ and

$$\mathcal{S} = \{b \in [0,T] : \gamma|_{[0,b]} \text{ aims at } q\}.$$

We have just shown that $\varepsilon \in \mathcal{S}$. Let $A = \sup \mathcal{S} > 0$. By continuity of the distance function, it is easy to see that $\mathcal{S}$ is closed, and therefore $A \in \mathcal{S}$. If $A = T$, then $\gamma|_{[0,T]}$ is a geodesic of length $T = d(p,q)$ that aims at q, and by the remark above we are done. So we assume $A < T$ and derive a contradiction.

Let $y = \gamma(A)$, and choose $\delta > 0$ such that $\overline{B}_\delta(y)$ is a closed geodesic ball (Figure 6.16). The fact that $A \in \mathcal{S}$ means

$$d(y,q) = d(p,q) - d(p,y) = T - A.$$

Let $z \in S_\delta(y)$ be a point where $d(z,q)$ attains its minimum, and let $\tau\colon [0,\delta] \to M$ be the radial geodesic from y to z. By exactly the same argument as before, τ aims at q, so

$$d(z,q) = d(y,q) - d(y,z) = (T - A) - \delta. \tag{6.7}$$

By the triangle inequality and (6.7),

$$\begin{aligned} d(p,z) &\geq d(p,q) - d(z,q) \\ &= T - (T - A - \delta) = A + \delta. \end{aligned}$$

Therefore, the admissible curve consisting of $\gamma|_{[0,A]}$ (of length A) followed by τ (of length δ) is a minimizing curve from p to z. This means it has no corners, so z must lie on γ, and in fact $z = \gamma(A+\delta)$. But then (6.7) says

$$d(p,q) = T = (A + \delta) + d(z,q) = d(p,z) + d(z,q),$$

so $\gamma|_{[0,A+\delta]}$ aims at q and $A+\delta \in \mathcal{S}$, which is a contradiction. This completes the proof that there is a minimizing geodesic from p to q.

Finally, we need to show that Cauchy sequences converge. Let $\{q_i\}$ be a Cauchy sequence in M. For each i, let $\gamma_i(t) = \exp_p(tV_i)$ be a unit speed minimizing geodesic from p to q_i, and let $d_i = d(p,q_i)$, so that $q_i = \exp_p(d_i V_i)$

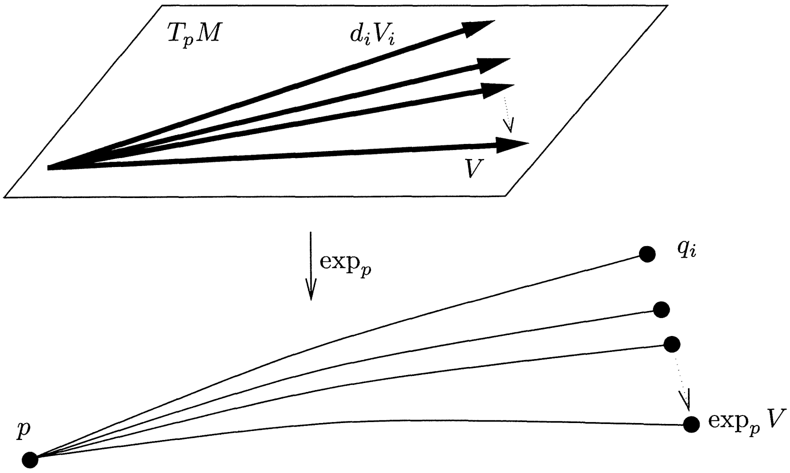


FIGURE 6.17. Cauchy sequences converge.

(Figure 6.17). The sequence $\{d_i\}$ is bounded in $\mathbf{R}$ (because Cauchy sequences in any metric space are bounded), and the sequence $\{V_i\}$ consists of unit vectors in T_pM, so the sequence of vectors $\{d_iV_i\}$ in T_pM is bounded. Therefore a subsequence $\{d_{i_k}V_{i_k}\}$ converges to $V \in T_pM$. By continuity of the exponential map, $q_{i_k} = \exp_p(d_{i_k}V_{i_k}) \to \exp_p V$, and since the original sequence $\{q_i\}$ is Cauchy, it converges to the same limit. This completes the proof of the Hopf–Rinow theorem. □

Because of this theorem, a connected Riemannian manifold is simply said to be *complete* if it is complete in either of the two equivalent senses discussed above. Complete manifolds are the natural setting for global questions in Riemannian geometry.

Exercise 6.5. Show that $\mathbf{R}^n$, $\mathbf{H}^n_R$, and $\mathbf{S}^n_R$ are complete.

We conclude this chapter by stating three important corollaries, whose proofs are immediate. The first two are corollaries of the proof of the Hopf–Rinow theorem, while the last one follows from its statement. In all of these corollaries, M is assumed to be a connected Riemannian manifold.

Corollary 6.14. *If there exists one point $p \in M$ such that the restricted exponential map $\exp_p$ is defined on all of T_pM, then M is complete.*

Corollary 6.15. *M is complete if and only if any two points in M can be joined by a minimizing geodesic segment.*

Corollary 6.16. *If M is compact, then every geodesic can be defined for all time.*

Problems

6-1. Define a connection on $\mathbf{R}^3$ by setting

$$\Gamma_{12}^3 = \Gamma_{23}^1 = \Gamma_{31}^2 = 1,$$
$$\Gamma_{21}^3 = \Gamma_{32}^1 = \Gamma_{13}^2 = -1,$$

and all other Christoffel symbols to zero. Show that this connection is compatible with the Euclidean metric and has minimizing geodesics, but is not symmetric.

6-2. We now have two kinds of "metrics" on a Riemannian manifold—the Riemannian metric and the distance function. Correspondingly, there are two definitions of "isometry" between Riemannian manifolds—a *Riemannian isometry* is a diffeomorphism that pulls one Riemannian metric back to the other, and a *metric isometry* is a homeomorphism that pulls one distance function back to the other. Prove that these two kinds of isometry are identical. [Hint: For the hard direction, first use the exponential map to show the homeomorphism is smooth.]

6-3. Suppose M and $\widetilde{M}$ are Riemannian manifolds (not necessarily complete), and $\varphi_i \colon M \to \widetilde{M}$ are Riemannian isometries that converge uniformly to a map $\varphi \colon M \to \widetilde{M}$. (This means that for any $\varepsilon > 0$, there exists I such that $d(\varphi_i(p), \varphi(p)) < \varepsilon$ for all $p \in M$ and all $i \geq I$.) Show that φ is a Riemannian isometry.

6-4. A subset $\mathcal{U}$ of a Riemannian manifold M is said to be *convex* if for each $p, q \in \mathcal{U}$, there is a unique (in M) minimizing geodesic from p to q lying entirely in $\mathcal{U}$. Show that every point has a convex neighborhood, as follows:

(a) Let $p \in M$ be fixed, and let $\mathcal{W}$ be a uniformly normal neighborhood of p. For $\varepsilon > 0$ small enough that $B_{2\varepsilon}(p) \subset \mathcal{W}$, define a subset $\mathcal{W}_\varepsilon \subset TM \times \mathbf{R}$ by

$$\mathcal{W}_\varepsilon = \{(q, V, t) \in TM \times \mathbf{R} : \\ q \in B_\varepsilon(p), V \in T_qM, |V| = 1, |t| < 2\varepsilon\}.$$

Define $f \colon \mathcal{W}_\varepsilon \to \mathbf{R}$ by

$$f(q, V, t) = d(\exp_q(tV), p)^2.$$

Show that f is smooth. [Hint: Use normal coordinates centered at p.]

(b) Show that if ε is chosen small enough, then $\partial^2 f / \partial t^2 > 0$ on $\mathcal{W}_\varepsilon$. [Hint: Compute $f(p, V, t)$ explicitly and use continuity.]

(c) If $q_1, q_2 \in B_\varepsilon(p)$ and γ is a minimizing geodesic from q_1 to q_2, show that $d(\gamma(t), p)$ attains its maximum at one of the endpoints of γ.

(d) Show that $B_\varepsilon(p)$ is convex.

6-5. If M is a complete Riemannian manifold and $N \subset M$ is a closed, embedded submanifold with the induced Riemannian metric, show that N is complete. [*Warning:* The distance function on N induced from the metric space structure of M is *not* in general equal to the Riemannian distance function of N.]

6-6. A curve $\gamma\colon [0, b) \to M$ $(0 < b \leq \infty)$ is said to *converge to infinity* if for every compact set $K \subset M$, there is a time $T \in [0, b)$ such that $\gamma(t) \notin K$ for $t > T$. (This means that γ converges to the "point at infinity" in the one-point compactification of M.) Prove that a Riemannian manifold is complete if and only if every regular curve that converges to infinity has infinite length. (The length of a curve whose domain is not compact is just the supremum of the lengths of its restrictions to compact subintervals.)

6-7. Show that any homogeneous Riemannian manifold is complete.

6-8. Suppose M is a complete Riemannian manifold that is isotropic at each point (see page 33). Show that M is homogeneous. [Hint: Given $p, q \in M$, consider the midpoint of a geodesic joining p and q.]

6-9. Generalize the first variation formula (Lemma 6.5) to the case of a variation that is not proper.

6-10. Let N be a closed, embedded submanifold of a Riemannian manifold M. For any point $p \in M - N$, we define the *distance from p to N* to be

$$d(p, N) := \inf\{d(p, x) : x \in N\}.$$

If $q \in N$ is a point such that $d(p, q) = d(p, N)$, and γ is any minimizing geodesic from p to q, prove that γ intersects N orthogonally. [Hint: Use Problem 6-9.]

6-11. Suppose $\widetilde{M}$ and M are Riemannian manifolds, and $p\colon \widetilde{M} \to M$ is a smooth covering map that is also a local isometry. If either M or $\widetilde{M}$ is complete, show that the other is also.

7 Curvature

In this chapter, we begin our study of the local invariants of Riemannian metrics. Starting with the question of whether all Riemannian metrics are locally isometric, we are led to a definition of the Riemannian curvature tensor as a measure of the failure of second covariant derivatives to commute. Then we prove the main result of this chapter: A manifold has zero curvature if and only if it is flat, that is, locally isometric to Euclidean space. At the end of the chapter, we derive the basic symmetries of the curvature tensor, and introduce the Ricci and scalar curvatures. The results of this chapter apply essentially unchanged to pseudo-Riemannian metrics.

Local Invariants

An important question about Riemannian manifolds is the following: Are they all locally isometric (i.e., given Riemannian n-manifolds $M, \widetilde{M}$ and points $p \in M$ and $\tilde{p} \in \widetilde{M}$, is there necessarily an isometry from a neighborhood of p to a neighborhood of $\tilde{p}$)? Or are there nontrivial local invariants that must be preserved by isometries? This is not an idle question, since many interesting and useful structures in differential geometry do *not* have local invariants. Some examples are as follows:

- *Nonvanishing vector fields.* In suitable coordinates, every nonvanishing vector field can be written locally as $V = \partial/\partial x^1$, so they are all locally equivalent.

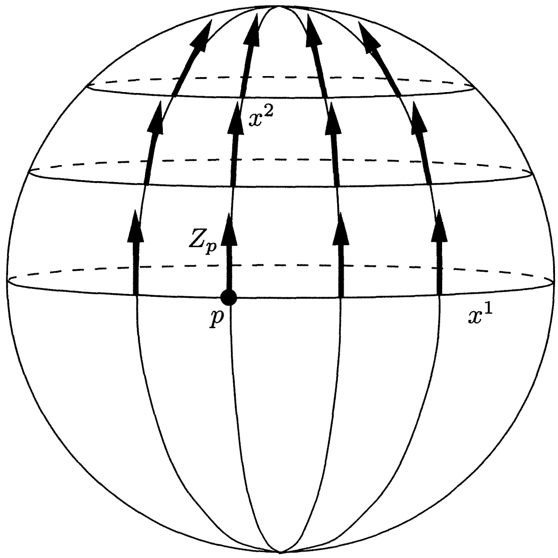


FIGURE 7.1. Result of parallel translation along the x^1-axis and the x^2-coordinate lines.

- *Riemannian metrics on a 1-manifold.* If $\gamma\colon I \to M$ is a local unit speed parametrization of a Riemannian 1-manifold, then $s = \gamma^{-1}$ gives a coordinate chart in which the metric has the expression $g = ds^2$. Thus every Riemannian 1-manifold is locally isometric to $\mathbf{R}$.

- *Symplectic forms.* A *symplectic form* is a closed 2-form ω that is nondegenerate, i.e., $\omega(X, Y) = 0$ for all $Y \in T_pM$ only if $X = 0$. The theorem of Darboux states that every symplectic form can be written in suitable coordinates as $\sum dx^i \wedge dy^i$. Thus all symplectic forms on $2n$-manifolds are locally equivalent.

On the other hand, you have shown in Problem 5-4 that the round 2-sphere and the Euclidean plane are not locally isometric. The key idea of that problem is that every tangent vector in the plane can be extended to a parallel vector field, so any Riemannian manifold that is locally isometric to $\mathbf{R}^2$ must have the same property locally.

Given a Riemannian 2-manifold M, there is an obvious way to attempt to construct such an extension of a vector $Z_p \in T_pM$. Choose any local coordinates (x^1, x^2) centered at p; first parallel translate Z_p along the x^1-axis, and then parallel translate the resulting vectors along the coordinate lines parallel to the x^2-axis (Figure 7.1). The result is a vector field Z that, by construction, is parallel along every x^2-coordinate line and along the x^1-axis. The question is whether this vector field is parallel along x^1-

coordinate lines other than the x^1-axis itself, or in other words, whether $\nabla_{\partial_1} Z \equiv 0$. Observe that $\nabla_{\partial_1} Z$ vanishes when $x^2 = 0$, so by uniqueness of parallel translates it would suffice to show that

$$\nabla_{\partial_2} \nabla_{\partial_1} Z = 0. \tag{7.1}$$

If we knew that

$$\nabla_{\partial_2} \nabla_{\partial_1} Z = \nabla_{\partial_1} \nabla_{\partial_2} Z, \tag{7.2}$$

then (7.1) would follow immediately because $\nabla_{\partial_2} Z = 0$ everywhere by construction. Indeed, on $\mathbf{R}^2$, direct computation shows

$$\begin{aligned} \nabla_{\partial_2} \nabla_{\partial_1} Z &= \nabla_{\partial_2} \left(\partial_1 Z^k \partial_k \right) \\ &= \partial_2 \partial_1 Z^k \partial_k, \end{aligned}$$

and $\nabla_{\partial_1} \nabla_{\partial_2} Z$ is equal to the same thing, because ordinary second partial derivatives commute. However, (7.2) might not hold for an arbitrary Riemannian metric; indeed, it is precisely the noncommutativity of such second covariant derivatives that forces this construction to fail on the sphere. Lurking behind this noncommutativity is the fact that the sphere is "curved."

To express this noncommutativity in a coordinate-invariant way, let's look more closely at the quantity $\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z$. On the Euclidean plane, we just showed that this always vanishes if $X = \partial_1$ and $Y = \partial_2$; however, for arbitrary vector fields this may no longer be true. In fact, in $\mathbf{R}^n$ with the Euclidean metric we have

$$\nabla_X \nabla_Y Z = \nabla_X \left(Y Z^k \partial_k \right) = X Y Z^k \partial_k,$$

and similarly $\nabla_Y \nabla_X Z = Y X Z^k \partial_k$. The difference between these two expressions is $(XYZ^k - YXZ^k)\partial_k = \nabla_{[X,Y]} Z$. Therefore the following relation holds for all vector fields X, Y, Z on $\mathbf{R}^n$:

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z. \tag{7.3}$$

By naturality of the Riemannian connection, it must also hold on any Riemannian manifold that is locally isometric to $\mathbf{R}^n$. We'll call (7.3) the *flatness criterion.*

This motivates the following definition. If M is any Riemannian manifold, the *(Riemann) curvature endomorphism* is the map $R\colon \mathcal{T}(M) \times \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$ defined by

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z.$$

Proposition 7.1. *The curvature endomorphism is a $\binom{3}{1}$-tensor field.*

Proof. By the tensor characterization lemma, we need only show that R is multilinear over $C^\infty(M)$. It is obviously multilinear over $\mathbf{R}$. For $f \in C^\infty(M)$,

$$\begin{aligned} R(X,fY)Z &= \nabla_X\nabla_{fY}Z - \nabla_{fY}\nabla_X Z - \nabla_{[X,fY]}Z \\ &= \nabla_X(f\nabla_Y Z) - f\nabla_Y\nabla_X Z - \nabla_{f[X,Y]+(Xf)Y}Z \\ &= (Xf)\nabla_Y Z + f\nabla_X\nabla_Y Z - f\nabla_Y\nabla_X Z \\ &\quad - f\nabla_{[X,Y]}Z - (Xf)\nabla_Y Z \\ &= fR(X,Y)Z. \end{aligned}$$

The same proof shows that R is linear over $C^\infty(M)$ in X, because $R(X,Y)Z = -R(Y,X)Z$ from the definition. The remaining case to be checked is linearity over $C^\infty(M)$ in Z; this is left to the reader. □

Exercise 7.1. Prove that $R(X,Y)(fZ) = fR(X,Y)Z$.

As a $\binom{3}{1}$-tensor field, the curvature endomorphism can be written in terms of any local frame with one upper and three lower indices. We adopt the convention that the *last* index is the contravariant (upper) one. (This is contrary to our default assumption that contravariant indices come first.) Thus, for example, the curvature endomorphism can be written in terms of local coordinates (x^i) as

$$R = R_{ijk}{}^l dx^i \otimes dx^j \otimes dx^k \otimes \partial_l,$$

where the coefficients $R_{ijk}{}^l$ are defined by

$$R(\partial_i, \partial_j)\,\partial_k = R_{ijk}{}^l \partial_l.$$

We also define the (*Riemann*) *curvature tensor* as the covariant 4-tensor field $Rm = R^\flat$ obtained from the $\binom{3}{1}$-tensor field R by lowering the last index. Its action on vector fields is given by

$$Rm(X,Y,Z,W) = \langle R(X,Y)Z, W\rangle, \tag{7.4}$$

and in coordinates it is written

$$Rm = R_{ijkl} dx^i \otimes dx^j \otimes dx^k \otimes dx^l,$$

where $R_{ijkl} = g_{lm}R_{ijk}{}^m$.

It is appropriate to note here that there is much variation in the literature with respect to the sign conventions adopted in the definitions of the Riemann curvature endomorphism and curvature tensor. While almost all authors define the curvature endomorphism as we have, there are a few (notably [dC92, GHL87]) whose definition is the negative of ours. There is much less agreement on the sign of the curvature tensor: whichever sign is

chosen for the curvature endomorphism, you will see the curvature tensor defined as in (7.4) but with various permutations of (X, Y, Z, W) on the right-hand side. After applying the symmetries of the curvature tensor that we will prove at the end of this chapter, however, all the definitions agree up to sign. There are various arguments to support one choice or another; we have made a choice that makes equation (7.4) easy to remember. You just have to be careful when you begin reading any book or article to determine the author's sign convention.

One reason the curvature endomorphism and curvature tensor are interesting is shown by the following lemma.

Lemma 7.2. *The Riemann curvature endomorphism and curvature tensor are local isometry invariants. More precisely, if* $\varphi\colon (M, g) \to (\widetilde{M}, \tilde{g})$ *is a local isometry, then*

$$\varphi^* \widetilde{Rm} = Rm;$$
$$\widetilde{R}(\varphi_* X, \varphi_* Y)\varphi_* Z = \varphi_*(R(X,Y)Z).$$

Exercise 7.2. Prove Lemma 7.2.

Flat Manifolds

To give a qualitative geometric meaning to the curvature tensor, we will show that it is precisely the obstruction to being locally isometric to Euclidean space. (In Chapter 8, after we have developed more machinery, we will be able to give a far more detailed quantitative interpretation.)

A Riemannian manifold is said to be *flat* if it is locally isometric to Euclidean space, that is, if every point has a neighborhood that is isometric to an open set in $\mathbf{R}^n$ with its Euclidean metric.

Theorem 7.3. *A Riemannian manifold is flat if and only if its curvature tensor vanishes identically.*

Proof. One direction is immediate: we showed above that the Euclidean metric satisfies the flatness criterion (7.3). Thus its curvature endomorphism is identically zero, and hence so also is its curvature tensor. If (M, g) is flat, in a neighborhood of any point there is an isometry φ to an open set in $(\mathbf{R}^n, \bar{g})$, and Lemma 7.2 shows that the curvature tensor of g is the pullback of that of $\bar{g}$, and thus is zero.

Now suppose (M, g) has vanishing curvature tensor. This means that the curvature endomorphism vanishes as well, so the flatness criterion (7.3) holds for all vector fields on M. We begin by showing that g shares one important property with the Euclidean metric: g admits a parallel orthonormal frame in a neighborhood of any point.

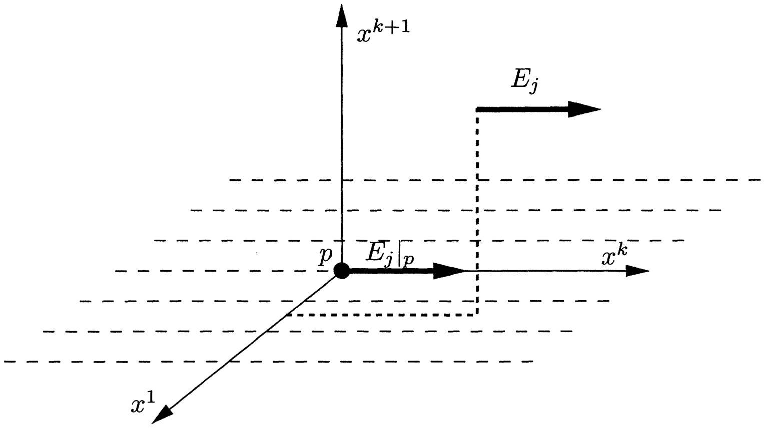


FIGURE 7.2. Proof that zero curvature implies flatness.

Let $p \in M$, and choose any orthonormal basis $(E_1|_p, \dots, E_n|_p)$ for T_pM. Let (x^i) be any coordinates centered at p such that $E_i|_p = \partial_i$ (for example, normal coordinates would suffice). By shrinking the coordinate neighborhood if necessary, we may assume that the image of the coordinate chart is a cube $\mathcal{C}_\varepsilon = \{x : |x^i| < \varepsilon,\ i = 1, \dots, n\}$.

Begin by parallel translating each vector $E_j|_p$ along the x^1-axis; then from each point on the x^1-axis, parallel translate along the coordinate line parallel to the x^2-axis; then successively parallel translate along coordinate lines parallel to the x^3 through x^n-axes (Figure 7.2). The result is n vector fields $(E_1, \dots, E_n)$ defined in $\mathcal{C}_\varepsilon$. The fact that the resulting vector fields are smooth follows from an inductive application of the theorem concerning smooth dependence of solutions to ODEs on initial conditions [Boo86, Theorem IV.4.2]; the details are left to the reader.

Because parallel translation preserves inner products, it is easy to see that the vector fields $\{E_j\}$ form an orthonormal frame. Since $\nabla_X E_j$ is linear over $C^\infty(M)$ in X, to show that the frame is parallel it suffices to show that $\nabla_{\partial_i} E_j = 0$ for each i and j.

Fix j. By construction, $\nabla_{\partial_1} E_j = 0$ on the x^1-axis, $\nabla_{\partial_2} E_j = 0$ on the (x^1, x^2)-plane, and in general $\nabla_{\partial_k} E_j = 0$ on the slice $M_k \subset \mathcal{C}_\varepsilon$ defined by $x^{k+1} = \cdots = x^m = 0$. We prove the following fact by induction on k:

$$\nabla_{\partial_1} E_j = \cdots = \nabla_{\partial_k} E_j = 0 \quad \text{on } M_k. \tag{7.5}$$

For $k = 1$, this is true by construction, and for $k = n$, it means that E_j is parallel on the whole cube $\mathcal{C}_\varepsilon$. So assume that (7.5) holds for some k. On M_{k+1}, $\nabla_{\partial_{k+1}} E_j = 0$ by construction, and for $i \le k$, $\nabla_{\partial_i} E_j = 0$ on the hyperplane where $x^{k+1} = 0$ by the inductive hypothesis. So it suffices to show that $\nabla_{\partial_{k+1}}(\nabla_{\partial_i} E_j) \equiv 0$. Since $[\partial_{k+1}, \partial_i] = 0$, the flatness criterion

gives

$$\nabla_{\partial_{k+1}}(\nabla_{\partial_i} E_j) = \nabla_{\partial_i}(\nabla_{\partial_{k+1}} E_j) = 0,$$

which completes the inductive step to show that the E_js are parallel.

Because the Riemannian connection is symmetric, we have

$$[E_i, E_j] = \nabla_{E_i} E_j - \nabla_{E_j} E_i = 0.$$

Thus the vector fields $(E_1, \ldots, E_n)$ form a commuting orthonormal frame on $\mathcal{C}_\varepsilon$. An important fact from elementary differential geometry is the following "normal form for commuting vector fields": *If $(E_1, \ldots, E_n)$ are commuting independent vector fields on a neighborhood of $p \in M$, there are coordinates (y^i) on a (possibly smaller) neighborhood of p such that $E_i = \partial/\partial y^i$.* (See [Boo86, p. 161], where the proof of this normal form is a key step in the proof of the Frobenius theorem.) In any such coordinates, $g_{ij} = g(\partial_i, \partial_j) = g(E_i, E_j) = \delta_{ij}$, so the map $y = (y^1, \ldots, y^n)$ is an isometry from a neighborhood of p to an open subset of Euclidean space. □

Exercise 7.3. Prove that the vector fields $\{E_j\}$ constructed in the preceding proof are smooth.

Exercise 7.4. Prove (or look up) the normal form theorem used in the preceding proof.

Symmetries of the Curvature Tensor

The curvature tensor on a Riemannian manifold has a number of symmetries besides the obvious skew-symmetry in its first two arguments.

Proposition 7.4. (Symmetries of the Curvature Tensor) *The curvature tensor has the following symmetries for any vector fields W, X, Y, Z:*

(a) $Rm(W,X,Y,Z) = -Rm(X,W,Y,Z)$.

(b) $Rm(W,X,Y,Z) = -Rm(W,X,Z,Y)$.

(c) $Rm(W,X,Y,Z) = Rm(Y,Z,W,X)$.

(d) $Rm(W,X,Y,Z) + Rm(X,Y,W,Z) + Rm(Y,W,X,Z) = 0$.

Before we begin the proof, a few remarks are in order. First, as the proof will show, (a) is a trivial consequence of the definition of the curvature endomorphism; (b) follows from the compatibility of the Riemannian connection with the metric; (d) follows from the symmetry of the connection; and (c) follows from (a), (b), and (d). The symmetry expressed in (d) is

called the *algebraic Bianchi identity* (or, more traditionally but less informatively, the *first Bianchi identity*). It is easy to show using (a)–(d) that a three-term sum obtained by cyclically permuting *any* three indices of Rm is also zero. Finally, it is useful to record the form of these symmetries in terms of components with respect to any basis:

(a′) $R_{ijkl} = -R_{jikl}$.

(b′) $R_{ijkl} = -R_{ijlk}$.

(c′) $R_{ijkl} = R_{klij}$.

(d′) $R_{ijkl} + R_{jkil} + R_{kijl} = 0$.

Proof of Proposition 7.4. Identity (a) is immediate from the obvious fact that $R(W,X)Y = -R(X,W)Y$. To prove (b), it suffices to show that $Rm(W,X,Y,Y) = 0$ for all Y, for then (b) follows from the expansion of $Rm(W,X,Y+Z,Y+Z) = 0$. Using compatibility with the metric, we have

$$\begin{aligned} WX|Y|^2 &= W(2\langle\nabla_X Y, Y\rangle) = 2\langle\nabla_W\nabla_X Y, Y\rangle + 2\langle\nabla_X Y, \nabla_W Y\rangle; \\ XW|Y|^2 &= X(2\langle\nabla_W Y, Y\rangle) = 2\langle\nabla_X\nabla_W Y, Y\rangle + 2\langle\nabla_W Y, \nabla_X Y\rangle; \\ [W,X]|Y|^2 &= 2\langle\nabla_{[W,X]}Y, Y\rangle. \end{aligned}$$

When we subtract the second and third equations from the first, the left-hand side is zero. The terms $2\langle\nabla_X Y, \nabla_W Y\rangle$ and $2\langle\nabla_W Y, \nabla_X Y\rangle$ cancel on the right-hand side, giving

$$\begin{aligned} 0 &= 2\langle\nabla_W\nabla_X Y, Y\rangle - 2\langle\nabla_X\nabla_W Y, Y\rangle - 2\langle\nabla_{[W,X]}Y, Y\rangle \\ &= 2\langle R(W,X)Y, Y\rangle \\ &= 2Rm(W,X,Y,Y). \end{aligned}$$

Next we prove (d). From the definition of Rm, this will follow immediately from

$$R(W,X)Y + R(X,Y)W + R(Y,W)X = 0.$$

Using the definition of R and the symmetry of the connection, the left-hand side expands to

$$\begin{aligned} &(\nabla_W\nabla_X Y - \nabla_X\nabla_W Y - \nabla_{[W,X]}Y) \\ &\quad + (\nabla_X\nabla_Y W - \nabla_Y\nabla_X W - \nabla_{[X,Y]}W) \\ &\quad + (\nabla_Y\nabla_W X - \nabla_W\nabla_Y X - \nabla_{[Y,W]}X) \\ &\quad = \nabla_W(\nabla_X Y - \nabla_Y X) + \nabla_X(\nabla_Y W - \nabla_W Y) + \nabla_Y(\nabla_W X - \nabla_X W) \\ &\qquad - \nabla_{[W,X]}Y - \nabla_{[X,Y]}W - \nabla_{[Y,W]}X \\ &\quad = \nabla_W[X,Y] + \nabla_X[Y,W] + \nabla_Y[W,X] \\ &\qquad - \nabla_{[W,X]}Y - \nabla_{[X,Y]}W - \nabla_{[Y,W]}X \\ &\quad = [W,[X,Y]] + [X,[Y,W]] + [Y,[W,X]]. \end{aligned}$$

This is zero by the Jacobi identity.

Finally, we show that identity (c) follows from the other three. Writing the algebraic Bianchi identity four times with indices cyclically permuted gives

$$\begin{aligned}
Rm(W,X,Y,Z) + Rm(X,Y,W,Z) + Rm(Y,W,X,Z) &= 0\\
Rm(X,Y,Z,W) + Rm(Y,Z,X,W) + Rm(Z,X,Y,W) &= 0\\
Rm(Y,Z,W,X) + Rm(Z,W,Y,X) + Rm(W,Y,Z,X) &= 0\\
Rm(Z,W,X,Y) + Rm(W,X,Z,Y) + Rm(X,Z,W,Y) &= 0.
\end{aligned}$$

Now add up all four equations. Applying (b) four times makes all the terms in the first two columns cancel. Then applying (a) and (b) in the last column yields $2Rm(Y,W,X,Z) - 2Rm(X,Z,Y,W) = 0$, which is equivalent to (c). □

There is one more identity that is satisfied by the covariant derivatives of the curvature tensor on any Riemannian manifold. Classically, it is called the *second Bianchi identity,* but modern authors tend to use the more informative name *differential Bianchi identity.*

Proposition 7.5. (Differential Bianchi Identity) *The total covariant derivative of the curvature tensor satisfies the following identity:*

$$\nabla Rm(X,Y,Z,V,W) + \nabla Rm(X,Y,V,W,Z) + \nabla Rm(X,Y,W,Z,V) = 0. \tag{7.6}$$

In components, this is

$$R_{ijkl;m} + R_{ijlm;k} + R_{ijmk;l} = 0. \tag{7.7}$$

Proof. First of all, by the symmetries of Rm, (7.6) is equivalent to

$$\nabla Rm(Z,V,X,Y,W) + \nabla Rm(V,W,X,Y,Z) + \nabla Rm(W,Z,X,Y,V) = 0. \tag{7.8}$$

This can be proved by a long and tedious computation, but there is a standard shortcut for such calculations in Riemannian geometry that makes our task immeasurably easier. To prove (7.6) holds at a particular point p, by multilinearity it suffices to prove the formula when X,Y,Z,V,W are basis elements with respect to some frame. The shortcut consists of choosing a special frame for each point p to simplify the computations there.

Let (x^i) be normal coordinates at p, and let X,Y,Z,V,W be arbitrary coordinate basis vectors ∂_i. These vectors satisfy two properties that simplify our computations enormously: (1) their commutators vanish identically, since $[\partial_i, \partial_j] \equiv 0$; and (2) their covariant derivatives vanish at p, since $\Gamma_{ij}^k(p) = 0$ (Proposition 5.11(f)).

Using these facts and the compatibility of the connection with the metric, the first term in (7.8) evaluated at p becomes

$$\begin{aligned}\nabla_W Rm(Z,V,X,Y) &= \nabla_W\langle R(Z,V)X,Y\rangle\\ &= \langle \nabla_W\nabla_Z\nabla_V X - \nabla_W\nabla_V\nabla_Z X, Y\rangle.\end{aligned}$$

Write this equation three times, with the vector fields W, Z, V cyclically permuted. Summing all three gives

$$\begin{aligned}\nabla Rm&(Z,V,X,Y,W) + \nabla Rm(V,W,X,Y,Z) + \nabla Rm(W,Z,X,Y,V)\\ &= \langle \nabla_W\nabla_Z\nabla_V X - \nabla_W\nabla_V\nabla_Z X\\ &\qquad + \nabla_Z\nabla_V\nabla_W X - \nabla_Z\nabla_W\nabla_V X\\ &\qquad + \nabla_V\nabla_W\nabla_Z X - \nabla_V\nabla_Z\nabla_W X, Y\rangle\\ &= \langle R(W,Z)\nabla_V X + R(Z,V)\nabla_W X + R(V,W)\nabla_Z X, Y\rangle\\ &= 0,\end{aligned}$$

where the last line follows because $\nabla_V X = \nabla_W X = \nabla_Z X = 0$ at p. □

Ricci and Scalar Curvatures

Because 4-tensors are so complicated, it is often useful to construct simpler tensors that summarize some of the information contained in the curvature tensor. The most important such tensor is the *Ricci curvature* or *Ricci tensor,* denoted Rc (or often Ric in the literature), which is the covariant 2-tensor field defined as the trace of the curvature endomorphism on its first and last indices. The components of Rc are usually denoted R_{ij}, so that

$$R_{ij} := R_{kij}{}^k = g^{km}R_{kijm}.$$

The *scalar curvature* is the function S defined as the trace of the Ricci tensor:

$$S := \operatorname{tr}_g Rc = R_i{}^i = g^{ij}R_{ij}.$$

Lemma 7.6. *The Ricci curvature is a symmetric 2-tensor field. It can be expressed in any of the following ways:*

$$R_{ij} = R_{kij}{}^k = R_{ik}{}^k{}_j = -R_{ki}{}^k{}_j = -R_{ikj}{}^k.$$

Exercise 7.5. Prove Lemma 7.6, using the symmetries of the curvature tensor.

Lemma 7.7. (Contracted Bianchi Identity) *The covariant derivatives of the Ricci and scalar curvatures satisfy the following identity:*

$$\operatorname{div} Rc = \frac{1}{2}\nabla S,$$

where div *is the divergence operator (Problem 3-3). In components, this is*

$$R_{ij;}{}^{j} = \frac{1}{2}S_{;i}. \tag{7.9}$$

Proof. Formula (7.9) follows immediately by contracting the component form (7.7) of the differential Bianchi identity on the indices i, l and then again on j, k, after raising one index of each pair. □

It is important to note that if the sign convention chosen for the curvature tensor is the opposite of ours, then the Ricci tensor must be defined as the trace of Rm on the first and third (or second and fourth) indices. (Of course the trace on the first two or last two indices is always zero by antisymmetry.) The definition is chosen so that the Ricci and scalar curvatures have the same meaning for everyone, regardless of the conventions chosen for the full curvature tensor. So, for example, if a manifold is said to have positive scalar curvature, there is no ambiguity as to what is meant.

A Riemannian metric is said to be an *Einstein metric* if its Ricci tensor is a scalar multiple of the metric at each point—that is, for some function λ, $Rc = \lambda g$ everywhere. Taking traces of both sides and noting that

$$\operatorname{tr}_g g = g_{ij}g^{ji} = \delta_i^i = \dim M,$$

we find that $\lambda = \frac{1}{n}S$ (where $n = \dim M$). Thus the Einstein condition can be written

$$Rc = \frac{1}{n}Sg. \tag{7.10}$$

Proposition 7.8. *If g is an Einstein metric on a connected manifold of dimension $n \geq 3$, its scalar curvature is constant.*

Proof. Taking the covariant derivative of each side of (7.10) and noting that the covariant derivative of the metric is zero, we see that the Einstein condition implies

$$R_{ij;k} = \frac{1}{n}S_{;k}g_{ij}.$$

Tracing this equation on j and k, and comparing with the contracted Bianchi identity (7.9), we conclude

$$\frac{1}{2}S_{;i} = \frac{1}{n}S_{;i}.$$

When $n > 2$, this implies $S_{;i} = 0$. But $S_{;i}$ is the component of $\nabla S = ds$, so connectedness of M implies S is constant. □

By an argument analogous to those of Chapter 5, Hilbert showed (see [Bes87, Theorem 4.21]) that Einstein metrics are critical points for the *total scalar curvature functional* $\mathcal{S}(g) := \int_M S\,dV$ on the space of all metrics on M with fixed volume. Thus Einstein metrics can be viewed as "optimal" metrics in a certain sense, and as such they form an appealing higher-dimensional analogue of the metrics of constant Gaussian curvature on 2-manifolds, with which one might hope to prove some sort of generalization of the uniformization theorem (Theorem 1.7 of Chapter 1). Although the statement of such a theorem cannot be as elegant as that of its 2-dimensional ancestor because there are known examples of smooth, compact manifolds that admit no Einstein metrics [Bes87, chapter 6], there is still a reasonable hope that "most" higher-dimensional manifolds (in some sense) admit Einstein metrics. This is an active and wide-open field of current research. See [Bes87] for a sweeping survey of recent research on Einstein metrics.

The term "Einstein metric" originated, as you might guess, in physics: The central assertion of Einstein's general theory of relativity is that physical space-time is modeled by a 4-manifold that carries a Lorentz metric whose Ricci curvature satisfies the following *Einstein field equation:*

$$Rc - \frac{1}{2}Sg = T, \tag{7.11}$$

where T is a certain symmetric 2-tensor (the *stress-energy tensor*) that describes the density, momentum, and stress of the matter and energy present at each point in space-time. It is shown in physics books (e.g. [HE73]) that (7.11) is the variational equation of a certain functional, called the *Hilbert action,* on the space of all Lorentz metrics on a given 4-manifold. Einstein's theory can then be interpreted as the assertion that a physically realistic space-time must be a critical point for this functional.

In the special case when $T \equiv 0$, (7.11) reduces to the *vacuum Einstein field equation* $Rc = \frac{1}{2}Sg$. Taking traces of both sides and recalling that $\operatorname{tr}_g g = \dim M = 4$, we obtain $S = 2S$, which implies $S = 0$. Therefore the vacuum Einstein equation is equivalent to $Rc = 0$, which means that g is a (pseudo-Riemannian) Einstein metric in the mathematical sense of the word. (At one point in the development of the theory, Einstein considered adding a term λg to the left-hand side of (7.11), where λ is a constant that he called the *cosmological constant.* With this modification the vacuum Einstein field equation would be exactly the same as the mathematicians' Einstein equation. Einstein soon decided, however, that the cosmological constant was a mistake on physical grounds.)

Other than these special cases and the obvious formal analogy between (7.11) and (7.10), there is no direct connection between the physicists' version of the Einstein equation and the mathematicians' version. Mathematically, Einstein metrics are interesting not because of their relation

to physics, but because of their potential applications to uniformization in higher dimensions.

Another approach to generalizing the uniformization theorem to higher dimensions is to search for metrics of constant scalar curvature. These are also critical points of the total scalar curvature functional, but only with respect to variations of the metric within a given conformal equivalence class. Thus it makes sense to ask whether, given a metric g on a manifold M, there exists a metric $\tilde{g}$ conformal to g that has constant scalar curvature. This is called the *Yamabe problem,* because it was first posed in 1960 by Hidehiko Yamabe, who claimed to have proved that the answer is always "yes" when M is compact. Yamabe's proof was later found to be in error, and it was two dozen years before the proof was finally completed by Richard Schoen; see [LP87] for an expository account of Schoen's solution. When M is noncompact, the issues are much subtler, and much current research is focused on determining exactly which conformal classes contain metrics of constant scalar curvature.

Problems

7-1. Let M be a Riemannian manifold, and (x^i) any local coordinates on M.

(a) Compute the components of the Riemann curvature tensor in terms of the Christoffel symbols in coordinates.

(b) Now suppose (x^i) are normal coordinates centered at $p \in M$. Show that the following holds at p:

$$R_{ijkl} = \frac{1}{2}\left(\partial_j \partial_l g_{ik} + \partial_i \partial_k g_{jl} - \partial_i \partial_l g_{jk} - \partial_j \partial_k g_{il}\right).$$

7-2. Let ∇ be the Riemannian connection on a Riemannian manifold (M, g), and let $\omega_i{}^j$ be its connection 1-forms with respect to a local frame $\{E_i\}$ (Problem 4-5). Define a matrix of 2-forms $\Omega_i{}^j$, called the *curvature 2-forms,* by

$$\Omega_i{}^j = \frac{1}{2} R_{kli}{}^j \varphi^k \wedge \varphi^l.$$

Show that they satisfy *Cartan's second structural equation:*

$$\Omega_i{}^j = d\omega_i{}^j - \omega_i{}^k \wedge \omega_k{}^j.$$

[Hint: Expand $R(E_k, E_l)E_i$ in terms of ∇ and $\omega_i{}^j$.]

7-3. If η is a 1-form, let

$$\eta_{i;jk} dx^i \otimes dx^j \otimes dx^k$$

be the local expression for $\nabla^2 \eta$. Prove the *Ricci identity*

$$\eta_{i;jk} - \eta_{i;kj} = R_{jki}{}^l \eta_l.$$

[Hint: Instead of expanding out the components of $\eta_{i;jk}$ in terms of the Christoffel symbols, either try to find an expression for $\nabla^2\eta$ similar to (4.8) and use the definition of the curvature endomorphism, or use the result of Problem 7-2.]

7-4. Let ∇ be any linear connection on a manifold M. We can define the curvature endomorphism of ∇ by the same formula as in the Riemannian case; ∇ is said to be flat if $R(X, Y)Z \equiv 0$. Prove that the following are equivalent:

(a) ∇ is flat.

(b) Near every point $p \in M$, there exists a parallel local frame.

(c) For all $p, q \in M$, parallel translation along a curve segment γ from p to q depends only on the homotopy class of γ.

(d) Parallel translation around any sufficiently small closed curve is the identity; that is, for any $p \in M$, there exists a neighborhood $\mathcal{U}$ of p such that if $\gamma\colon [a,b] \to \mathcal{U}$ is a smooth curve in $\mathcal{U}$ starting and ending at p, then $P_{ab}\colon T_pM \to T_pM$ is the identity map.

7-5. Let G be a Lie group with a bi-invariant metric g (see Problems 3-12 and 5-11). Show that the Riemannian curvature endomorphism of g can be computed as follows:

$$R(X,Y)Z = \frac{1}{4}[Z,[X,Y]]$$

whenever X, Y, Z are left-invariant vector fields on G.

8

Riemannian Submanifolds

This chapter has a dual purpose: first to develop the basic concepts of the theory of Riemannian submanifolds, and then to use these concepts to derive a quantitative interpretation of the curvature tensor.

After introducing some basic definitions and terminology concerning submanifolds, we define a tensor field called the second fundamental form, which measures the way a submanifold curves within the ambient manifold. We then prove the fundamental relationships between the intrinsic and extrinsic geometries of a submanifold: the Gauss formula relates the Riemannian connection on the submanifold to that of the ambient manifold, and the Gauss equation relates their curvatures. We show how the second fundamental form can be interpreted as a measure of the extrinsic curvature of submanifold geodesics.

Using these tools, we focus on the special case of hypersurfaces in $\mathbf{R}^{n+1}$, and show how the second fundamental form is related to the principal curvatures and Gaussian curvature. We prove Gauss's *Theorema Egregium*, which shows that the Gaussian curvature of a surface in $\mathbf{R}^3$ can be computed from the intrinsic curvature tensor.

In the last section, we introduce the promised quantitative geometric interpretation of the curvature tensor. It allows us to compute sectional curvatures, which are just the Gaussian curvatures of 2-dimensional submanifolds swept out by geodesics tangent to 2-planes in the tangent space. Finally, we compute the sectional curvatures of our model Riemannian manifolds—Euclidean spaces, spheres, and hyperbolic spaces.

Caution must be exercised when applying the methods of this chapter to pseudo-Riemannian manifolds, because the restriction of a pseudo-

Riemannian metric to a submanifold might not be nondegenerate. (See Problem 8-9, though.)

Riemannian Submanifolds and the Second Fundamental Form

Definitions

Suppose $(\widetilde{M}, \tilde{g})$ is a Riemannian manifold of dimension m, M is a manifold of dimension n, and $\iota\colon M \to \widetilde{M}$ is an immersion. If M is given the induced Riemannian metric $g := \iota^*\tilde{g}$, then ι is said to be an *isometric immersion* (or an *isometric embedding* if ι happens to be an embedding). If in addition ι is injective, so that M is an (immersed or embedded) submanifold of $\widetilde{M}$, then M is said to be a *Riemannian submanifold* of $\widetilde{M}$. In all of these situations, $\widetilde{M}$ is called the *ambient manifold.*

All the considerations of this chapter apply to any isometric immersion. Since our computations are all local, and since any immersion is locally an embedding, we may assume M is an embedded Riemannian submanifold, possibly after shrinking M a bit. We usually proceed under such an assumption without further comment. Covariant derivatives and curvatures with respect to (M, g) are written in the normal way, while those with respect to $(\widetilde{M}, \tilde{g})$ are written with tildes. We can unambiguously use the inner-product notation $\langle X, Y\rangle$ to refer to either metric g or $\tilde{g}$, since g is just the restriction of $\tilde{g}$ to TM.

It is easy to see that the set

$$T\widetilde{M}|_M := \coprod_{p\in M} T_p\widetilde{M}$$

is a smooth vector bundle over M, with local trivializations provided, for example, by the vector fields $(\partial_1, \dots, \partial_m)$ in any coordinate chart on $\widetilde{M}$. We call it the *ambient tangent bundle over M.* Any smooth vector field on $\widetilde{M}$ clearly restricts to a smooth section of $T\widetilde{M}|_M$. Conversely, any smooth section X of $T\widetilde{M}|_M$ can be extended to a smooth section of $T\widetilde{M}$ by the same method as in the proof of Exercise 2.3. When there is no risk of confusion, we use the same letter to denote both a vector field or function on M and its extension to $\widetilde{M}$.

At each $p \in M$, the ambient tangent space $T_p\widetilde{M}$ splits as an orthogonal direct sum $T_p\widetilde{M} = T_pM \oplus N_pM$, where $N_pM := (T_pM)^\perp$ is the *normal space* at p with respect to the inner product $\tilde{g}$ on $T_p\widetilde{M}$ (Figure 8.1). The set

$$NM := \coprod_{p\in M} N_pM$$

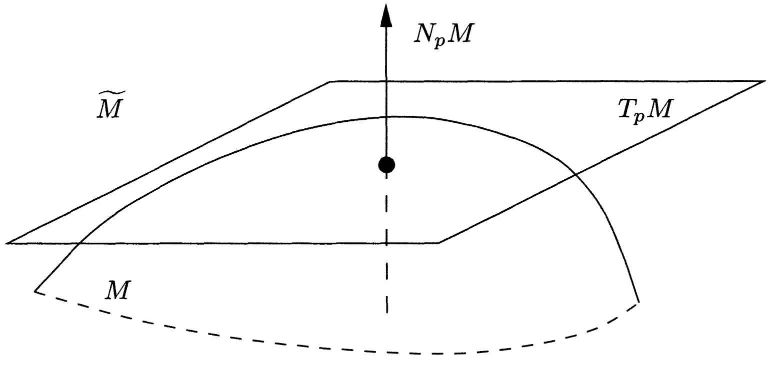


FIGURE 8.1. The normal space at p.

is called the *normal bundle* of M. To see that it is a smooth vector bundle over M, we use the result of Problem 3-1: Given any point $p \in M$, there is a neighborhood $\widetilde{\mathcal{U}}$ of p in $\widetilde{M}$ and a smooth orthonormal frame $(E_1, \dots, E_m)$ on $\widetilde{\mathcal{U}}$, called an *adapted orthonormal frame,* such that the restrictions of $(E_1, \dots, E_n)$ to M form a local orthonormal frame for TM. Given any such frame, the last $m - n$ vectors $(E_{n+1}|_p, \dots, E_m|_p)$ form a basis for N_pM at each $p \in M$, and we can use the components of a normal vector with respect to this basis to construct a local trivialization of NM. It is straightforward to check that the transition functions are smooth, so NM is a vector bundle by Lemma 2.2. The notations $\mathcal{T}(\widetilde{M}|_M)$ and $\mathcal{N}(M)$ denote the spaces of smooth sections of $T\widetilde{M}|_M$ and NM, respectively.

Projecting orthogonally at each point $p \in M$ onto the subspaces T_pM and N_pM gives maps called the *tangential* and *normal projections*

$$\begin{aligned} \pi^\top &: T\widetilde{M}|_M \to TM \\ \pi^\perp &: T\widetilde{M}|_M \to NM. \end{aligned}$$

In terms of an adapted orthonormal frame, these are just the usual projections onto $\operatorname{span}(E_1, \dots, E_n)$ and $\operatorname{span}(E_{n+1}, \dots, E_m)$ respectively, so both projections map smooth sections to smooth sections. If X is a section of $T\widetilde{M}|_M$, we often use the shorthand notations $X^\top := \pi^\top X$ and $X^\perp := \pi^\perp X$ for its tangential and normal projections.

The Second Fundamental Form

Our first main task is to compare the Riemannian connection of M with that of $\widetilde{M}$. The starting point for doing so is the orthogonal decomposition of sections of $T\widetilde{M}|_M$ into tangential and orthogonal components as above. If X, Y are vector fields in $\mathcal{T}(M)$, we can extend them to vector fields on $\widetilde{M}$,

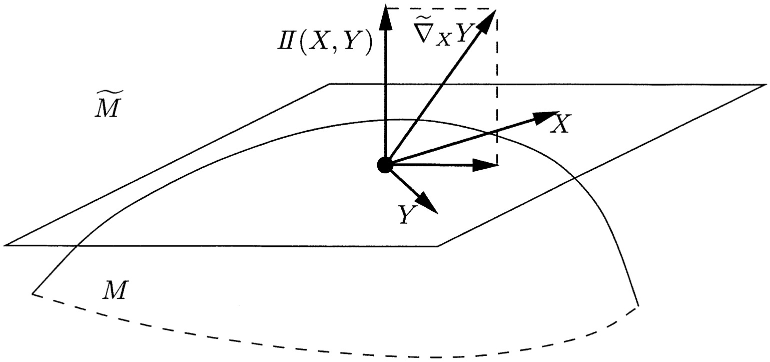


FIGURE 8.2. The second fundamental form.

apply the ambient covariant derivative operator $\widetilde{\nabla}$, and then decompose at points of M to get

$$\widetilde{\nabla}_X Y = (\widetilde{\nabla}_X Y)^\top + (\widetilde{\nabla}_X Y)^\perp. \tag{8.1}$$

We would like to interpret the two terms on the right-hand side of this decomposition.

Let's focus first on the normal component. We define the *second fundamental form* of M to be the map $I\!I$ (read "two") from $\mathcal{T}(M) \times \mathcal{T}(M)$ to $\mathcal{N}(M)$ given by

$$I\!I(X,Y) := (\widetilde{\nabla}_X Y)^\perp,$$

where X and Y are extended arbitrarily to $\widetilde{M}$ (Figure 8.2). Since $\pi^\perp$ maps smooth sections to smooth sections, $I\!I(X,Y)$ is a smooth section of NM.

The term "first fundamental form," by the way, was used classically to refer to the induced metric g on M. Although that usage has mostly been replaced by more descriptive terminology, we seem unfortunately to be stuck with the name "second fundamental form." The word "form" in both cases refers to bilinear form, not differential form.

Lemma 8.1. *The second fundamental form is*

(*a*) *independent of the extensions of X and Y;*

(*b*) *bilinear over $C^\infty(M)$; and*

(*c*) *symmetric in X and Y.*

Proof. First we show that the symmetry of $I\!I$ follows from the symmetry of the connection $\widetilde{\nabla}$. Let X and Y be extended arbitrarily to M. Then

$$I\!I(X,Y) - I\!I(Y,X) = (\widetilde{\nabla}_X Y - \widetilde{\nabla}_Y X)^\perp = [X,Y]^\perp.$$

Since X and Y are tangent to M at all points of M, so is their Lie bracket. (This follows easily from Exercise 2.3.) Therefore $[X,Y]^\perp = 0$, so $I\!I$ is symmetric.

Because $\widetilde{\nabla}_X Y|_p$ depends only on X_p, it is clear that $I\!I(X,Y)$ is independent of the extension chosen for X, and that $I\!I(X,Y)$ is linear over $C^\infty(M)$ in X. By symmetry, the same is true for Y. □

We have not yet identified the tangential term in the decomposition of $\widetilde{\nabla}_X Y$. The following theorem shows that it is nothing other than $\nabla_X Y$, the covariant derivative with respect to the Riemannian connection of g. Therefore, we can interpret the second fundamental form as a measure of the difference between the intrinsic Riemannian connection on M and the ambient Riemannian connection on $\widetilde{M}$.

Theorem 8.2. (The Gauss Formula) *If $X, Y \in \mathcal{T}(M)$ are extended arbitrarily to vector fields on $\widetilde{M}$, the following formula holds along M:*

$$\widetilde{\nabla}_X Y = \nabla_X Y + I\!I(X,Y).$$

Proof. Because of the decomposition (8.1) and the definition of the second fundamental form, it suffices to show that $(\widetilde{\nabla}_X Y)^\top = \nabla_X Y$ at all points of M.

Define a map $\nabla^\top \colon \mathcal{T}(M) \times \mathcal{T}(M) \to \mathcal{T}(M)$ by

$$\nabla^\top_X Y := (\widetilde{\nabla}_X Y)^\top,$$

where X, Y are extended arbitrarily to $\widetilde{M}$. We examined a special case of this construction, in which $\tilde{g}$ is the Euclidean metric, in Lemma 5.1. It follows exactly as in the proof of that lemma that $\nabla^\top$ is a connection on M. Once we show that it is symmetric and compatible with g, the uniqueness of the Riemannian connection on M shows that $\nabla^\top = \nabla$.

To see that $\nabla^\top$ is symmetric, we use the symmetry of $\widetilde{\nabla}$ and the fact that $[X,Y]$ is tangent to M:

$$\begin{aligned}\nabla^\top_X Y - \nabla^\top_Y X &= (\widetilde{\nabla}_X Y - \widetilde{\nabla}_Y X)^\top \\ &= [X,Y]^\top = [X,Y].\end{aligned}$$

To prove compatibility with g, let $X, Y, Z \in \mathcal{T}(M)$ be extended arbitrarily to $\widetilde{M}$. Using compatibility of $\widetilde{\nabla}$ with $\tilde{g}$, and evaluating at points of M,

$$\begin{aligned}X\langle Y,Z\rangle &= \langle \widetilde{\nabla}_X Y, Z\rangle + \langle Y, \widetilde{\nabla}_X Z\rangle \\ &= \langle (\widetilde{\nabla}_X Y)^\top, Z\rangle + \langle Y, (\widetilde{\nabla}_X Z)^\top\rangle \\ &= \langle \nabla^\top_X Y, Z\rangle + \langle Y, \nabla^\top_X Z\rangle.\end{aligned}$$

Therefore $\nabla^\top$ is compatible with g, so $\nabla^\top = \nabla$. □

Although the second fundamental form is defined in terms of covariant derivatives of vector fields *tangent* to M, it can also be used to evaluate covariant derivatives of *normal* vector fields, as the following lemma shows.

Lemma 8.3. (The Weingarten Equation) *Suppose $X, Y \in \mathfrak{T}(M)$ and $N \in \mathcal{N}(M)$. When X, Y, N are extended arbitrarily to $\widetilde{M}$, the following equation holds at points of M:*

$$\langle \widetilde{\nabla}_X N, Y \rangle = -\langle N, I\!I(X,Y) \rangle.$$

Proof. Since $\langle N, Y \rangle$ vanishes identically along M and X is tangent to M, the following holds along M:

$$\begin{aligned} 0 &= X\langle N, Y \rangle \\ &= \langle \widetilde{\nabla}_X N, Y \rangle + \langle N, \widetilde{\nabla}_X Y \rangle \\ &= \langle \widetilde{\nabla}_X N, Y \rangle + \langle N, \nabla_X Y + I\!I(X,Y) \rangle \\ &= \langle \widetilde{\nabla}_X N, Y \rangle + \langle N, I\!I(X,Y) \rangle. \end{aligned}$$

□

In addition to describing the difference between the intrinsic and extrinsic connections, the second fundamental form plays an even more important role in describing the difference between the curvature tensors of $\widetilde{M}$ and M. The explicit formula, also due to Gauss, is given in the following theorem.

Theorem 8.4. (The Gauss Equation) *For any $X, Y, Z, W \in T_pM$, the following equation holds:*

$$\begin{aligned} \widetilde{Rm}(X,Y,Z,W) = Rm&(X,Y,Z,W) \\ &- \langle I\!I(X,W), I\!I(Y,Z) \rangle + \langle I\!I(X,Z), I\!I(Y,W) \rangle. \end{aligned}$$

Proof. Let X, Y, Z, W be extended arbitrarily to vector fields on M, and then to vector fields on $\widetilde{M}$ that are tangent to M at points of M. Along M, the Gauss formula gives

$$\begin{aligned} \widetilde{Rm}(X,Y,Z,W) &= \left\langle \widetilde{\nabla}_X \widetilde{\nabla}_Y Z - \widetilde{\nabla}_Y \widetilde{\nabla}_X Z - \widetilde{\nabla}_{[X,Y]} Z, W \right\rangle \\ &= \Big\langle \widetilde{\nabla}_X (\nabla_Y Z + I\!I(Y,Z)) - \widetilde{\nabla}_Y (\nabla_X Z + I\!I(X,Z)) \\ &\qquad - (\nabla_{[X,Y]} Z + I\!I([X,Y],Z)), W \Big\rangle. \end{aligned}$$

Since the second fundamental form takes its values in the normal bundle and W is tangent to M, the last $I\!I$ term is zero. Apply the Weingarten equation to the other two terms involving $I\!I$ (with $I\!I(Y,Z)$ or $I\!I(X,Z)$

playing the role of N) to get

$$\begin{aligned}\widetilde{Rm}(X,Y,Z,W) &= \langle \widetilde{\nabla}_X \nabla_Y Z, W\rangle - \langle I\!I(Y,Z), I\!I(X,W)\rangle \\ &\quad - \langle \widetilde{\nabla}_Y \nabla_X Z, W\rangle + \langle I\!I(X,Z), I\!I(Y,W)\rangle \\ &\quad - \langle \nabla_{[X,Y]} Z, W\rangle.\end{aligned}$$

Decomposing each term involving $\widetilde{\nabla}$ into its tangential and normal components, we see that only the tangential component survives. Using the Gauss formula allows each to be rewritten in terms of ∇, giving

$$\begin{aligned}\widetilde{Rm}(X,Y,Z,W) &= \langle \nabla_X \nabla_Y Z, W\rangle - \langle \nabla_Y \nabla_X Z, W\rangle - \langle \nabla_{[X,Y]} Z, W\rangle \\ &\quad - \langle I\!I(Y,Z), I\!I(X,W)\rangle + \langle I\!I(X,Z), I\!I(Y,W)\rangle \\ &= \langle R(X,Y)Z, W\rangle \\ &\quad - \langle I\!I(X,W), I\!I(Y,Z)\rangle + \langle I\!I(X,Z), I\!I(Y,W)\rangle.\end{aligned}$$

This proves the theorem. □

Curvature of Curves

By studying the curvature of curves in Riemannian manifolds, we can give a more geometric interpretation to the second fundamental form. If $\gamma\colon I \to M$ is a unit speed curve in a Riemannian manifold, we define the (*geodesic*) *curvature* of γ as the function $\kappa\colon I \to \mathbf{R}$ given by

$$\kappa(t) = |D_t\dot\gamma(t)|.$$

If γ is an arbitrary regular curve (not necessarily unit speed), we first reparametrize it by arc length to get a unit speed curve, and then define the curvature by this formula as a function of arc length. Clearly κ vanishes identically if and only if γ is a geodesic, so it may be thought of as a quantitative measure of how far γ deviates from being a geodesic. If $M = \mathbf{R}^n$ with the Euclidean metric, this is the same as the classical notion of curvature introduced in advanced calculus courses.

Exercise 8.1. If γ is a unit speed curve in $\mathbf{R}^n$ and $\kappa(t_0) \neq 0$, show that there is a unique unit speed parametrized circle $c\colon \mathbf{R} \to \mathbf{R}^n$, called the *osculating circle* at $\gamma(t_0)$, with the property that c and γ have the same position, velocity, and acceleration at $t = t_0$. Show that $\kappa(t_0) = 1/R$, where R is the radius of the osculating circle.

Exercise 8.2. Suppose $\gamma\colon I \to M$ is a regular curve in a Riemannian manifold, but not necessarily unit speed. Show that the curvature of γ at $\gamma(t)$ is

$$\kappa = \frac{|D_t\dot\gamma(t)|}{|\dot\gamma(t)|^2} - \frac{\langle D_t\dot\gamma(t), \dot\gamma(t)\rangle}{|\dot\gamma(t)|^3}.$$

If $M \hookrightarrow \widetilde{M}$ is a Riemannian submanifold and γ is a curve in M, γ has two distinct geodesic curvatures: its "intrinsic" curvature κ as a curve in M, and its "extrinsic" curvature $\tilde{\kappa}$ as a curve in $\widetilde{M}$. The second fundamental form can be used to compute the relationship between the two. First we need another version of the Gauss formula, better suited to covariant derivatives along curves.

Lemma 8.5. (The Gauss Formula Along a Curve) *Let M be a Riemannian submanifold of $\widetilde{M}$, and γ a curve in M. For any vector field V tangent to M along γ,*

$$\widetilde{D}_t V = D_t V + I\!I(\dot{\gamma}, V).$$

Proof. In terms of an adapted orthonormal frame, V can be written $V(t) = V^i(t)E_i$, where the sum is only over $i = 1, \ldots, n$. Applying the product rule and the Gauss formula, we get

$$\begin{aligned}
\widetilde{D}_t V &= \dot{V}^i E_i + V^i \widetilde{\nabla}_{\dot{\gamma}} E_i \\
&= \dot{V}^i E_i + V^i \nabla_{\dot{\gamma}} E_i + V^i\, I\!I(\dot{\gamma}, E_i) \\
&= D_t V + I\!I(\dot{\gamma}, V).
\end{aligned}$$

□

Applying this lemma to the special case in which $V = \dot{\gamma}$, we obtain the following formula for the acceleration of any curve in M:

$$\widetilde{D}_t \dot{\gamma} = D_t \dot{\gamma} + I\!I(\dot{\gamma}, \dot{\gamma}).$$

If γ is a geodesic in M, this formula simplifies to

$$\widetilde{D}_t \dot{\gamma} = I\!I(\dot{\gamma}, \dot{\gamma}).$$

Thus we obtain the following concrete geometric interpretation of the second fundamental form: *For any vector $V \in T_pM$, $I\!I(V, V)$ is the $\tilde{g}$-acceleration at p of the g-geodesic γ_V. If V is a unit vector, $|I\!I(V, V)|$ is the $\tilde{g}$-curvature of γ_V at p.* Note that the second fundamental form is symmetric and bilinear, so it is completely determined by its values of the form $I\!I(V, V)$ as V ranges over unit vectors tangent to M.

In the special case in which $\widetilde{M}$ is $\mathbf{R}^m$ with the Euclidean metric, we can make this geometric interpretation even more concrete: *$I\!I(V, V)$ is the ordinary Euclidean acceleration of the geodesic in M with initial velocity V.*

Exercise 8.3. Suppose $M \subset \mathbf{R}^m$ is a submanifold with the induced Riemannian metric, γ is a curve in M, and V is a vector field tangent to M along γ.

(a) Show that $D_t V(t)$ is the orthogonal projection onto TM of the ordinary Euclidean derivative $\dot{V}(t)$.

(b) Show that γ is a geodesic in M if and only if its Euclidean acceleration $\ddot{\gamma}$ is everywhere normal to M.

(c) Use this to give another proof that the geodesics on the n-sphere are the great circles.

We say a Riemannian submanifold $M \subset \widetilde{M}$ is *totally geodesic* if for every $V \in TM$, the $\tilde{g}$-geodesic γ_V lies entirely in M.

Exercise 8.4. Show that the following are equivalent for a Riemannian submanifold $M \subset \widetilde{M}$:

(a) M is totally geodesic.

(b) Every g-geodesic in M is also a $\tilde{g}$-geodesic in $\widetilde{M}$.

(c) The second fundamental form of M vanishes identically.

Hypersurfaces in Euclidean Space

Now we specialize the preceding considerations to the case in which M is a hypersurface (i.e., a submanifold of codimension 1) in $\mathbf{R}^{n+1}$ with the induced Riemannian metric. We denote the Euclidean metric as usual by $\bar{g}$. Covariant derivatives and curvatures associated with $\bar{g}$ will be indicated by a bar.

In this situation, at each point of M there are exactly two unit normal vectors. If M is orientable (which we may assume by passing to a subset of M), we can use an orientation to pick out a unique normal. The resulting vector field N is a smooth section of NM, as can be seen easily by noting that in terms of any local adapted orthonormal frame $(E_1, \dots, E_{n+1})$, it must be $N = \pm E_{n+1}$. We will address as we go along the question of how various quantities depend on the choice of normal vector field.

The Scalar Second Fundamental Form and the Shape Operator

Given a unit normal vector field N, we can replace the vector-valued second fundamental form $I\!I$ by a somewhat simpler scalar-valued form. The *scalar second fundamental form* h is the symmetric 2-tensor on M defined by

$$h(X,Y) = \langle I\!I(X,Y), N \rangle.$$

Since N is a unit vector spanning NM at each point, this is equivalent to

$$I\!I(X,Y) = h(X,Y)N.$$

Note that the sign of h depends on which unit normal is chosen, but h is otherwise independent of choices.

Raising one index of h, we get a tensor field $s \in \mathcal{T}^1_1(M)$, which can also be thought of as a field of endomorphisms of TM by Lemma 2.4, called the *shape operator* of M. It is characterized by

$$\langle X, sY \rangle = h(X, Y) \qquad \text{for all } X, Y \in \mathcal{T}(M).$$

Because h is symmetric, s is a selfadjoint endomorphism of TM, that is,

$$\langle sX, Y \rangle = \langle X, sY \rangle \qquad \text{for all } X, Y \in \mathcal{T}(M).$$

As with h, the sign of s depends on the choice of N.

In terms of the tensor fields h and s, the formulas of the last section can be rewritten somewhat more simply. First, we have the *Gauss formula for Euclidean hypersurfaces:*

$$\overline{\nabla}_X Y = \nabla_X Y + h(X, Y)N.$$

Second, the Weingarten equation can be written

$$\langle \overline{\nabla}_X N, Y \rangle = -h(X, Y) = -\langle sX, Y \rangle. \tag{8.2}$$

Since $\langle \overline{\nabla}_X N, N \rangle = \frac{1}{2}\overline{\nabla}_X |N|^2 = 0$, it follows that $\overline{\nabla}_X N$ is tangent to M, so (8.2) is equivalent to the *Weingarten equation for Euclidean hypersurfaces:*

$$\overline{\nabla}_X N = -sX. \tag{8.3}$$

Finally, since $\overline{Rm} = 0$ on $\mathbf{R}^{n+1}$, the *Gauss equation for Euclidean hypersurfaces* is

$$Rm(X, Y, Z, W) = h(X, W)h(Y, Z) - h(X, Z)h(Y, W). \tag{8.4}$$

If γ is a curve in M, its Euclidean acceleration vector can be decomposed into tangential and normal components in the usual way. By the Gauss formula, they are

$$\ddot{\gamma} = \overline{D}_t \dot{\gamma} = D_t \dot{\gamma} + h(\dot{\gamma}, \dot{\gamma})N.$$

If γ is a unit speed geodesic in M, its intrinsic acceleration $D_t\dot{\gamma}$ is zero. Its Euclidean acceleration therefore has only a normal component,

$$\ddot{\gamma} = h(\dot{\gamma}, \dot{\gamma})N,$$

and its Euclidean curvature is

$$\overline{\kappa} = |\ddot{\gamma}| = |h(\dot{\gamma}, \dot{\gamma})|.$$

Therefore $h(\dot{\gamma}, \dot{\gamma}) = \pm\overline{\kappa}$, with a positive sign if and only if $\ddot{\gamma}$ points in the same direction as N. This shows that the scalar second fundamental form has the following geometric interpretation: *For any unit vector $V \in T_pM$, $h(V, V)$ is the signed Euclidean curvature at p of the M-geodesic γ_V (Figure 8.3), with a positive sign if γ_V is curving toward N at p and a negative sign if it is curving away from N.*

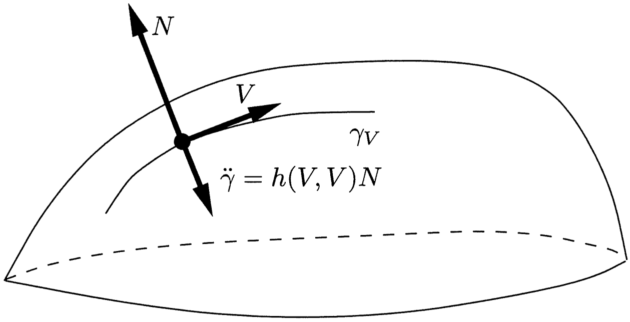


FIGURE 8.3. Geometric interpretation of $h(V, V)$.

Principal Curvatures

At any point $p \in M$, we have seen that the shape operator s is a selfadjoint linear transformation on the tangent space T_pM. From elementary linear algebra, any such operator has real eigenvalues $\kappa_1, \ldots, \kappa_n$, and there is an orthonormal basis $(E_1, \ldots, E_n)$ for T_pM consisting of s-eigenvectors, so that $sE_i = \kappa_i E_i$ (no summation). In this basis both h and s are diagonal, and h has the expression

$$h(X, Y) = \kappa_1 X^1 Y^1 + \cdots + \kappa_n X^n Y^n.$$

The eigenvalues of s are called the *principal curvatures* of M at p, and the corresponding eigenspaces are called the *principal directions.* They are independent of choice of basis, but the principal curvatures change sign if we change the normal vector. The principal curvatures give a concise description of the local shape of the embedded surface M, in a sense made precise by the following exercise.

Exercise 8.5. Suppose $M \subset \mathbf{R}^{n+1}$ is a hypersurface with the induced metric. Let $p \in M$, and let $\kappa_1, \ldots, \kappa_n$ denote the principal curvatures of M at p with respect to some choice of unit normal.

1. Show that M can be approximated locally by the quadratic polynomial $\frac{1}{2}h$, in the following sense: There are Euclidean coordinates $(x, y) = (x^1, \ldots, x^n, y)$ centered at p such that M is described locally by an equation of the form $y = f(x)$, where the second-order Taylor series of f at the origin is

$$f(x) = \frac{1}{2}(\kappa_1(x^1)^2 + \cdots + \kappa_n(x^n)^2) + O(|x|^3).$$

2. If $n = 2$, show that κ_1 and κ_2 are equal to the minimum and maximum signed Euclidean curvatures of M-geodesics passing through p, and also to the minimum and maximum signed Euclidean curvatures of plane curves obtained by intersecting M with planes orthogonal to T_pM.

Gaussian and Mean Curvatures

There are two combinations of the principal curvatures that play particularly important roles for Euclidean hypersurfaces. The *Gaussian curvature* is defined as $K = \det s$, and the *mean curvature* as $H = (1/n)\operatorname{tr} s = (1/n)\operatorname{tr}_g h$. Since the determinant and trace of a linear map are basis-independent, these are well defined. In terms of the principal curvatures, they are

$$K = \kappa_1\kappa_2\cdots\kappa_n; \qquad H = \frac{1}{n}(\kappa_1 + \cdots + \kappa_n).$$

We know from Proposition 5.13 that the geodesics on the round 2-sphere $\mathbf{S}^2_R$ of radius R are exactly the great circles of radius R. Since these have Euclidean curvature $1/R$, it is immediate that the principal curvatures at any point are $\kappa_1 = \kappa_2 = \pm 1/R$. Therefore S^2_R has constant mean curvature $H = \pm 1/R$ and constant Gaussian curvature $K = 1/R^2$.

For other surfaces in $\mathbf{R}^3$, the Gaussian and mean curvatures are usually easiest to compute in terms of parametrizations. Let $M \subset \mathbf{R}^3$ be a smooth surface, and let $X\colon \mathcal{U} \to \mathbf{R}^3$ be a local parametrization of M. The coordinates (u^1, u^2) on $\mathcal{U} \subset \mathbf{R}^2$ thus give local coordinates for M. The coordinate vector fields $\partial_i = \partial/\partial u^i$ push forward to vectors $X_*\partial_i = \partial_i X$ (thinking of $X(u) = (X^1(u), X^2(u), X^3(u))$ as a vector-valued function of u) in $\mathbf{R}^3$ that are tangent to M. Their ordinary cross product is therefore normal to M, so one choice of unit normal is

$$N = \frac{\partial_1 X \times \partial_2 X}{|\partial_1 X \times \partial_2 X|}.$$

We can then compute the shape operator using the Weingarten equation for Euclidean hypersurfaces (8.3):

$$s\partial_i = -\overline{\nabla}_{\partial_i} N = -\partial_i N,$$

where again we think of N as a vector-valued function of u, and use the fact that the directional derivative $\overline{\nabla}_{\partial_i}$ can be evaluated by differentiating along the u^i-coordinate curve in M. After expressing $s\partial_1$ and $s\partial_2$ in terms of the basis vectors $(\partial_1 X, \partial_2 X)$, it is straightforward to compute the principal curvatures. Problems 8-1, 8-2, and 8-3 will give you practice in carrying out these computations for surfaces presented in various ways.

A hypersurface M with mean curvature identically equal to zero is called *minimal*. The reason for this terminology is that, by an argument analogous to those of Chapter 6, one can show that $H \equiv 0$ is the variational equation for the surface area functional $\mathcal{A}(M) = \int_M dV$ (where dV is the Riemannian volume element for the induced metric g). Thus hypersurfaces with mean curvature zero are precisely the critical points for the functional $\mathcal{A}$. In particular, any hypersurface that minimizes surface area among those

with a fixed boundary has zero mean curvature, and a small enough piece of every minimal hypersurface is area minimizing. We do not pursue the subject any further in this book, but you can find a good introductory account in [Law80].

Clearly the mean curvature of any hypersurface changes sign if we change the sign of the normal vector field. If n is odd, the Gaussian curvature also changes sign, but if n is even (in particular for surfaces in $\mathbf{R}^3$), the Gaussian curvature is independent of the choice of N. In any case, both the Gaussian and mean curvatures are defined in terms of a particular embedding of M into $\mathbf{R}^{n+1}$, and there is little reason to suspect that they have much to do with the intrinsic Riemannian geometry of M with its induced metric g. The amazing discovery made by Gauss was that the Gaussian curvature of a surface in $\mathbf{R}^3$ is actually an *intrinsic* invariant of the Riemannian manifold (M, g).

Theorem 8.6. (Gauss's *Theorema Egregium*) *Let $M \subset \mathbf{R}^3$ be a 2-dimensional submanifold and g the induced metric on M. For any $p \in M$ and any basis (X, Y) for T_pM, the Gaussian curvature of M at p is given by*

$$K = \frac{Rm(X, Y, Y, X)}{|X|^2|Y|^2 - \langle X, Y\rangle^2}. \tag{8.5}$$

Therefore the Gaussian curvature is an isometry invariant of (M, g).

Proof. We begin with the special case in which $(X, Y) = (E_1, E_2)$ is an orthonormal basis for T_pM. In this case the denominator in (8.5) is equal to 1. If we write $h_{ij} = h(E_i, E_j)$, then in this basis $K = \det s = \det(h_{ij})$, and the Gauss equation (8.4) reads

$$Rm(E_1, E_2, E_2, E_1) = h_{11}h_{22} - h_{12}h_{21} = \det(h_{ij}) = K.$$

This is equivalent to (8.5).

Now let X, Y be any basis for T_pM. The Gram–Schmidt algorithm yields an orthonormal basis as follows:

$$E_1 = \frac{X}{|X|};$$

$$E_2 = \frac{Y - \langle Y, \frac{X}{|X|}\rangle \frac{X}{|X|}}{\left|Y - \langle Y, \frac{X}{|X|}\rangle \frac{X}{|X|}\right|} = \frac{Y - \frac{\langle Y, X\rangle}{|X|^2}X}{\left|Y - \frac{\langle Y, X\rangle}{|X|^2}X\right|}.$$

Then by the preceding computation, the Gaussian curvature at p is

$$\begin{aligned}
K &= Rm(E_1, E_2, E_2, E_1) \\
&= \frac{Rm\left(X, Y - \frac{\langle Y,X\rangle}{|X|^2}X, Y - \frac{\langle Y,X\rangle}{|X|^2}X, X\right)}{|X|^2\left|Y - \frac{\langle Y,X\rangle}{|X|^2}X\right|^2} \\
&= \frac{Rm(X,Y,Y,X)}{|X|^2\left(|Y|^2 - 2\frac{\langle Y,X\rangle^2}{|X|^2} + \frac{\langle Y,X\rangle^2}{|X|^2}\right)} \\
&= \frac{Rm(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X,Y\rangle^2}.
\end{aligned}$$

(In the third line, we used the fact that $Rm(X,X,\cdot,\cdot) = Rm(\cdot,\cdot,X,X) = 0$ by the symmetries of the curvature tensor.) This proves the theorem. □

Motivated by the *Theorema Egregium*, we define the *Gaussian curvature* K of an abstract Riemannian 2-manifold (M,g), not necessarily embedded in $\mathbf{R}^3$, by formula (8.5) in terms of any local frame (X,Y). In the special case in which M is a Riemannian submanifold of $\mathbf{R}^3$, the *Theorema Egregium* shows that this agrees with the extrinsic definition of K as the determinant of the scalar second fundamental form.

Lemma 8.7. *The Gaussian curvature of a Riemannian 2-manifold is related to the curvature tensor, Ricci tensor, and scalar curvature by the formulas*

$$\begin{aligned}
Rm(X,Y,Z,W) &= K\big(\langle X,W\rangle\langle Y,Z\rangle - \langle X,Z\rangle\langle Y,W\rangle\big); \\
Rc(X,Y) &= K\langle X,Y\rangle; \\
S &= 2K.
\end{aligned} \tag{8.6}$$

Thus K is independent of choice of frame, and completely determines the curvature tensor.

Proof. Since both sides of the first equation are tensors, we can compute them in terms of any basis. Let (E_1, E_2) be any orthonormal basis for T_pM, and consider the components $R_{ijkl} = Rm(E_i, E_j, E_k, E_l)$ of the curvature tensor. In terms of this basis, (8.5) gives $K = R_{1221}$. By antisymmetry, R_{ijkl} vanishes whenever $i = j$ or $k = l$, so the only nonzero components of Rm are

$$R_{1221} = R_{2112} = -R_{1212} = -R_{2121} = K.$$

Comparing $Rm(X,Y,Z,W)$ with $K(\langle X,W\rangle\langle Y,Z\rangle - \langle X,Z\rangle\langle Y,W\rangle)$ when each of X,Y,Z,W is either E_1 or E_2 proves the first equation of (8.6).

The components of the Ricci tensor in this basis are

$$R_{ij} = R_{1ij1} + R_{2ij2},$$

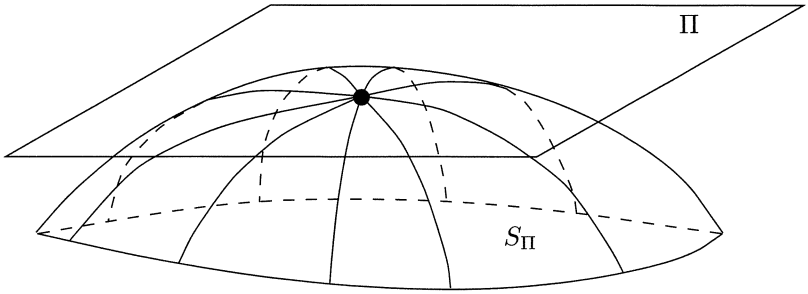


FIGURE 8.4. A plane section.

from which it follows easily that

$$R_{12} = R_{21} = 0; \qquad R_{11} = R_{22} = K,$$

which is equivalent to the second equation. Finally, the scalar curvature is

$$S = \operatorname{tr}_g Rc = R_{11} + R_{22} = 2K.$$

Because the scalar curvature is independent of choice of frame, so is K. □

Although the Ricci tensor always satisfies $Rc = Kg$ on a 2-manifold, this does not imply that K is constant in two dimensions, as you can see from the proof of Proposition 7.8. Thus the notion of an Einstein metric is not useful for 2-manifolds.

Exercise 8.6. Show that the hyperbolic plane $\mathbf{H}^2_R$ of radius R has constant Gaussian curvature $K = -1/R^2$. [Hint: Show that it suffices to compute K at one point; the coordinate computations are easiest at the origin in the disk model.]

Geometric Interpretation of Curvature in Higher Dimensions

Sectional Curvatures

Now we can give a quantitative geometric interpretation to the curvature tensor in any dimension. Let M be a Riemannian n-manifold and $p \in M$. If Π is any 2-dimensional subspace of T_pM, and $\mathcal{V} \subset T_pM$ is any neighborhood of zero on which $\exp_p$ is a diffeomorphism, then $S_\Pi := \exp_p(\Pi \cap \mathcal{V})$ is a 2-dimensional submanifold of M containing p (Figure 8.4), called the *plane section* determined by Π. Note that S_Π is just the set swept out by geodesics whose initial tangent vectors lie in Π.

We define the *sectional curvature* of M associated with Π, denoted $K(\Pi)$, to be the Gaussian curvature of the surface S_Π at p with the induced metric. If (X, Y) is any basis for Π, we also use the notation $K(X, Y)$ for $K(\Pi)$.

Proposition 8.8. *If (X, Y) is any basis for a 2-plane $\Pi \subset T_pM$, then*

$$K(X,Y) = \frac{Rm(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X, Y\rangle^2}. \tag{8.7}$$

Proof. For this proof, we denote the induced metric on S_Π by $\tilde{g}$, and continue to denote the metric on M by g. As in the first part of this chapter, we use tildes to denote geometric quantities associated with $\tilde{g}$, but note that now the roles of g and $\tilde{g}$ are reversed.

We claim first that the second fundamental form of S_Π vanishes at p. To see why, let $V \in \Pi \subset T_pM$, and let $\gamma = \gamma_V$ be the M-geodesic with initial velocity V, which lies in S_Π by definition. By the Gauss formula for vector fields along curves,

$$0 = D_t\dot{\gamma} = \widetilde{D}_t\dot{\gamma} + I\!I(\dot{\gamma}, \dot{\gamma}).$$

Since the two terms in this sum are orthogonal, each must vanish identically. Evaluating at $t = 0$ gives $I\!I(V, V) = 0$. Since V was an arbitrary element of T_pM and $I\!I$ is symmetric, this shows that $I\!I = 0$ at p. (We cannot in general expect $I\!I$ to vanish at other points of S_Π—it is only at p that all geodesics starting tangent to S remain in S.)

Now the Gauss equation tells us that the curvature tensors of S_Π and M are related at p by

$$\widetilde{Rm}(X,Y,Z,W) = Rm(X,Y,Z,W)$$

whenever $X, Y, Z, W \in \Pi$. In particular, the Gaussian curvature of S_Π at p is

$$K(\Pi) = \frac{\widetilde{Rm}(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X, Y\rangle^2} = \frac{Rm(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X, Y\rangle^2}.$$

This is what was to be proved. □

Thus one important class of quantitative information provided by the curvature tensor is the sectional curvatures of all plane sections. It turns out, in fact, that this is the *only* information contained in the curvature tensor: as the following lemma shows, the sectional curvatures completely determine the curvature tensor.

Lemma 8.9. *Suppose $\mathcal{R}_1$ and $\mathcal{R}_2$ are covariant 4-tensors on a vector space V with an inner product, and both have the symmetries of the curvature*

tensor (as described in Proposition 7.4). If for every pair of independent vectors $X, Y \in V$,

$$\frac{\mathcal{R}_1(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X,Y\rangle^2} = \frac{\mathcal{R}_2(X,Y,Y,X)}{|X|^2|Y|^2 - \langle X,Y\rangle^2},$$

then $\mathcal{R}_1 = \mathcal{R}_2$.

Proof. Setting $\mathcal{R} = \mathcal{R}_1 - \mathcal{R}_2$, it suffices to show $\mathcal{R} = 0$ under the assumption that $\mathcal{R}(X,Y,Y,X) = 0$ for all X, Y.

For any vectors X, Y, Z, since $\mathcal{R}$ also has the symmetries of the curvature tensor,

$$\begin{aligned} 0 &= \mathcal{R}(X+Y, Z, Z, X+Y) \\ &= \mathcal{R}(X,Z,Z,X) + \mathcal{R}(X,Z,Z,Y) + \mathcal{R}(Y,Z,Z,X) + \mathcal{R}(Y,Z,Z,Y) \\ &= 2\mathcal{R}(X,Z,Z,Y). \end{aligned}$$

From this it follows that

$$\begin{aligned} 0 &= \mathcal{R}(X, Z+W, Z+W, Y) \\ &= \mathcal{R}(X,Z,Z,Y) + \mathcal{R}(X,Z,W,Y) + \mathcal{R}(X,W,Z,Y) + \mathcal{R}(X,W,W,Y) \\ &= \mathcal{R}(X,Z,W,Y) + \mathcal{R}(X,W,Z,Y). \end{aligned}$$

Therefore $\mathcal{R}$ is antisymmetric in any adjacent pair of arguments. Now the algebraic Bianchi identity yields

$$\begin{aligned} 0 &= \mathcal{R}(X,Y,Z,W) + \mathcal{R}(Y,Z,X,W) + \mathcal{R}(Z,X,Y,W) \\ &= \mathcal{R}(X,Y,Z,W) - \mathcal{R}(Y,X,Z,W) - \mathcal{R}(X,Z,Y,W) \\ &= 3\mathcal{R}(X,Y,Z,W). \end{aligned}$$

□

We can also give a geometric interpretation for the Ricci and scalar curvatures. Given any unit vector $V \in T_pM$, choose an orthonormal basis $\{E_i\}$ for T_pM such that $E_1 = V$. Then $Rc(V,V)$ is given by

$$Rc(V,V) = R_{11} = R_{k11}{}^k = \sum_{k=1}^n Rm(E_k, E_1, E_1, E_k) = \sum_{k=2}^n K(E_1, E_k).$$

Therefore the Ricci tensor has the following interpretation: *For any unit vector $V \in T_pM$, $Rc(V,V)$ is the sum of the sectional curvatures of planes spanned by V and other elements of an orthonormal basis.* Since Rc is symmetric and bilinear, it is completely determined by its values of the form $Rc(V,V)$ for unit vectors V.

Similarly, the scalar curvature is

$$S = R_j{}^j = \sum_{j=1}^n Rc(E_j, E_j) = \sum_{j,k=1}^n Rm(E_k, E_j, E_j, E_k) = \sum_{j \neq k} K(E_j, E_k).$$

Therefore *the scalar curvature is the sum of all sectional curvatures of planes spanned by pairs of orthonormal basis elements.*

If the opposite sign convention is chosen for the curvature tensor, then the right-hand side of formula (8.7) has to be adjusted accordingly, with $Rm(X, Y, X, Y)$ taking the place of $Rm(X, Y, Y, X)$. This is so that whatever sign convention is chosen for the curvature tensor, the notion of positive or negative sectional curvature has the same meaning for everyone.

Sectional Curvatures of the Model Spaces

We can now compute the sectional curvatures of our three families of homogeneous model spaces. Note first that each model space has an isometry group that acts transitively on orthonormal frames, and so acts transitively on 2-planes in the tangent bundle. Therefore each has *constant sectional curvature,* which means that the sectional curvatures are the same for all planes at all points.

First we consider the simplest case: Euclidean space. Since the curvature tensor of $\mathbf{R}^n$ is identically zero, clearly all sectional curvatures are zero. This is obvious geometrically, since each plane section is actually a plane, which has zero Gaussian curvature.

Next consider the sphere $\mathbf{S}^n_R$. We need only compute the sectional curvature for the plane Π spanned by (∂_1, ∂_2) at the north pole. The geodesics with initial velocity in Π are great circles in the (x^1, x^2, x^{n+1}) subspace. Therefore S_Π is isometric to the round 2-sphere of radius R embedded in $\mathbf{R}^3$. As we showed earlier in this chapter, $\mathbf{S}^2_R$ has Gaussian curvature $1/R^2$. Therefore $\mathbf{S}^n_R$ has constant sectional curvature equal to $1/R^2$.

Finally we come to the hyperbolic spaces. It suffices to consider the point $N = (0, \dots, R)$ in the hyperboloid model, and the plane $\Pi \subset T_N \mathbf{H}^n_R$ spanned by $\partial/\partial\xi^1, \partial/\partial\xi^2$. The geodesics with initial velocities in Π are great hyperbolas lying in the (ξ^1, ξ^2, τ) subspace; they sweep out a 2-dimensional hyperboloid that is easily seen to be isometric to $\mathbf{H}^2_R$. By Exercise 8.6, therefore, $K(\Pi) = -1/R^2$, so H^n_R has constant sectional curvature $-1/R^2$. (See also Problem 8-9 for another approach.)

Exercise 8.7. Show that real projective space $\mathbf{RP}^n$ has a metric of constant positive sectional curvature.

Since the sectional curvatures determine the curvature tensor, one would expect to have an explicit formula for Rm when the sectional curvature is constant. Such a formula is provided in the following lemma.

Lemma 8.10. *Suppose (M, g) is any Riemannian n-manifold with constant sectional curvature C. The curvature endomorphism, curvature ten-*

sor, Ricci tensor, and scalar curvature of g are given by the formulas

$$\begin{aligned}
R(X,Y)Z &= C\big(\langle Y,Z\rangle X - \langle X,Z\rangle Y\big);\\
Rm(X,Y,Z,W) &= C\big(\langle X,W\rangle\langle Y,Z\rangle - \langle X,Z\rangle\langle Y,W\rangle\big);\\
Rc &= (n-1)Cg;\\
S &= n(n-1)C.
\end{aligned}$$

In terms of any basis,

$$\begin{aligned}
R_{ijkl} &= C(g_{il}g_{jk} - g_{ik}g_{jl});\\
R_{ij} &= (n-1)Cg_{ij}.
\end{aligned}$$

Exercise 8.8. Prove Lemma 8.10.

Problems

8-1. Let $M \subset \mathbf{R}^3$ be a surface of revolution as described in Exercise 3.3 and Problem 5-2.

(a) If the generating curve γ is unit speed, show that the Gaussian curvature of M is $-\ddot{a}(t)/a(t)$.

(b) Show that there is a surface of revolution in $\mathbf{R}^3$ that has constant Gaussian curvature equal to 1 but does not have constant mean curvature.

8-2. Suppose Ω is an open set in $\mathbf{R}^n$ and $f\colon \Omega \to \mathbf{R}$ is a smooth function. Let $M = \{(x, f(x)) : x \in \Omega\} \subset \mathbf{R}^{n+1}$ be the graph of f. Observe that the map $\varphi\colon \Omega \to M$ given by $\varphi(u) = (u, f(u))$ gives a global parametrization of M; the corresponding coordinates $(u^1, \ldots, u^n)$ on M are called *graph coordinates.*

(a) Let N be the upward-pointing unit normal vector field along M. Compute the components of the shape operator in graph coordinates, in terms of f and its partial derivatives.

(b) Let $M \subset \mathbf{R}^{n+1}$ be the paraboloid defined as the graph of $f(u) = |u|^2$. Compute the principal curvatures of M.

8-3. Let $\Omega \subset \mathbf{R}^{n+1}$ be an open set, $F\colon \Omega \to \mathbf{R}$ a smooth submersion, and $M = F^{-1}(0)$. (F is called a *defining function* for M.) Show that the scalar second fundamental form of M with respect to the unit normal vector field $N = \operatorname{grad} F/|\operatorname{grad} F|$ is given by

$$h(V, W) = -\frac{\partial_i \partial_j F V^i V^j}{|\operatorname{grad} F|^2},$$

where $V = V^i \partial_i$ in Euclidean coordinates on $\mathbf{R}^{n+1}$. Derive formulas for the Gaussian and mean curvatures of F in the case $n = 2$.

8-4. Let $M \subset \mathbf{R}^3$ be the *catenoid,* which is the surface of revolution obtained by revolving the curve $x = \cosh z$ around the z-axis. Show that M is a minimal surface.

8-5. Suppose $M \subset \widetilde{M}$ is a compact, embedded, Riemannian submanifold. For any $\varepsilon > 0$, let N_ε denote the subset $\{V : |V| < \varepsilon\}$ of the normal bundle NM, and M_ε the set of points in $\widetilde{M}$ whose Riemannian distance from M is less than ε.

(a) Prove the *tubular neighborhood theorem:* For ε sufficiently small, the restriction to N_ε of the exponential map of $\widetilde{M}$ is a diffeomorphism from N_ε to M_ε. Any open set M_ε that is the image of such a diffeomorphism is called a *tubular neighborhood* of M.

(b) If $r(x)$ denotes the distance from $x \in \widetilde{M}$ to M, show that r^2 is a smooth function on any tubular neighborhood M_ε. Give an example in which r^2 is not smooth on all of $\widetilde{M}$.

8-6. Let $M \subset \mathbf{R}^{n+1}$ be a hypersurface with the induced metric and N a smooth unit normal vector field along M. At each point $p \in M$, $N_p \in T_p\mathbf{R}^{n+1}$ can be thought of as a unit vector in $\mathbf{R}^{n+1}$ and therefore as a point in $\mathbf{S}^n$. Thus each choice of normal vector field defines a smooth map $N\colon M \to \mathbf{S}^n$, called the *Gauss map* of M. Show that $N^* d\mathring{V} = K\, dV_g$, where $d\mathring{V}$ is the volume element of $\mathbf{S}^n$ with the round metric, and K is the Gaussian curvature of M.

8-7. Suppose $g = g_1 \oplus g_2$ is a product metric on $M_1 \times M_2$ as in (3.3).

(a) Show that for each point $p_i \in M_i$, the submanifolds $M_1 \times \{p_2\}$ and $\{p_1\} \times M_2$ are totally geodesic.

(b) If $\Pi \subset T(M_1 \times M_2)$ is a 2-plane spanned by $X_1 \in TM_1$ and $X_2 \in TM_2$, show that $K(\Pi) = 0$.

(c) Show that the product metric on $S^2 \times S^2$ has nonnegative sectional curvature.

(d) Show that there is an embedding of T^2 in $S^2 \times S^2$ such that the induced metric is flat.

8-8. Consider the basis

$$X = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

for the Lie algebra $\mathfrak{su}(2)$. For each positive real number a, define a left-invariant metric g_a on the group $SU(2)$ by declaring aX, Y, Z to be an orthonormal frame. Compute the sectional curvatures with respect to g_a of the planes spanned by (X,Y), (Y,Z), and (Z,X). [Remark: $SU(2)$ is diffeomorphic to $\mathbf{S}^3$ by the map that sends $(\alpha,\beta) \in \mathbf{S}^3 \subset \mathbf{C}^2$ to $\begin{pmatrix} \alpha & \beta \\ -\bar\beta & \bar\alpha \end{pmatrix} \in SU(2)$. These metrics are called the *Berger metrics* on $\mathbf{S}^3$.]

8-9. This problem outlines another proof that the sectional curvature of $\mathbf{H}^n_R$ is $-1/R^2$.

(a) If $\widetilde{M}$ is a pseudo-Riemannian manifold, a submanifold $\iota\colon M \hookrightarrow \widetilde{M}$ is called *spacelike* if $\iota^* g$ is positive definite on M. Show that the Gauss formula and the Gauss equation hold for spacelike submanifolds.

(b) Prove that the sectional curvature of $\mathbf{H}^n_R$ is $-1/R^2$ by applying the Gauss equation to the hyperboloid model.

8-10. Suppose M is a connected n-dimensional Riemannian manifold, and a Lie group G acts effectively on M by isometries. (A group action is said to be *effective* if no element of G other than the identity acts as the identity on M.) Show that $\dim G \leq n(n+1)/2$, and that equality is possible only if M has constant sectional curvature.

8-11. Let $p\colon (\widetilde{M}, \tilde{g}) \to (M, g)$ be a Riemannian submersion (Problem 3-8). Using the notation and results of Problem 5-9, show that the sectional curvatures of g are related to those of $\tilde{g}$ by

$$K(X,Y) = \widetilde{K}(\widetilde{X},\widetilde{Y}) + \frac{3}{4}\left|[\widetilde{X},\widetilde{Y}]^V\right|^2,$$

for any pair X, Y of orthonormal vector fields on M.

8-12. Let $p\colon \mathbf{S}^{2n+1} \to \mathbf{CP}^n$ be the Riemannian submersion described in Problem 3-9. We identify $\mathbf{C}^{n+1}$ with $\mathbf{R}^{n+1} \times \mathbf{R}^{n+1}$ by means of coordinates $(x^1, \ldots, x^{n+1}, y^1, \ldots, y^{n+1})$ defined by $z^j = x^j + iy^j$.

(a) Show that the vector field

$$T = x^j \frac{\partial}{\partial y^j} - y^j \frac{\partial}{\partial x^j}$$

on $\mathbf{C}^{n+1}$ is tangent to $\mathbf{S}^{2n+1}$ and spans the vertical space V_z at each point $z \in \mathbf{S}^{2n+1}$.

(b) If W, Z are horizontal vector fields on $\mathbf{S}^{2n+1}$, show that

$$[W,Z]^V = -d\omega(W,Z)T = 2\langle W, JZ\rangle T,$$

where ω is the 1-form on $\mathbf{C}^{n+1}$ given by

$$\omega = T^\flat = \sum_j x^j\, dy^j - y^j\, dx^j,$$

and $J\colon T\mathbf{C}^{n+1} \to T\mathbf{C}^{n+1}$ is the orthogonal, real-linear map

$$J\left(X^j \frac{\partial}{\partial x^j} + Y^j \frac{\partial}{\partial y^j}\right) = X^j \frac{\partial}{\partial y^j} - Y^j \frac{\partial}{\partial x^j}.$$

(This is just multiplication by $i = \sqrt{-1}$ in complex coordinates.)

(c) Let W, Z be orthonormal vectors in $T\mathbf{CP}^n$. Show that the sectional curvature $K(W,Z)$ is

$$K(W,Z) = 1 + 3\langle \widetilde{W}, J\widetilde{Z}\rangle^2.$$

(See Problem 8-11.)

(d) If $n \geq 2$, show that at each point of $\mathbf{CP}^n$, the sectional curvatures take on all values between 1 and 4, inclusive. Compute the Gaussian curvature of $\mathbf{CP}^1$.

8-13. Suppose (M, g) is a 3-dimensional Riemannian manifold that is homogeneous and isotropic. Show that g has constant sectional curvature. Show that the analogous result in dimension 4 is not true. [Hint: See Problem 8-12.]

8-14. Let G be a Lie group with a bi-invariant metric g (see Problem 7-5).

(a) Show that the sectional curvatures of g are all nonnegative.

(b) If $H \subset G$ is a Lie subgroup, show that H is totally geodesic.

(c) If H is connected, show that it is flat in the induced metric if and only if it is Abelian.

9
The Gauss–Bonnet Theorem

We are finally in a position to prove our first major local-global theorem in Riemannian geometry: the Gauss–Bonnet theorem. This is a local-global theorem *par excellence*, because it asserts the equality of two very differently defined quantities on a compact, orientable Riemannian 2-manifold M: the integral of the Gaussian curvature, which is determined by the local geometry of M; and 2π times the Euler characteristic of M, which is a global topological invariant. Although it applies only in two dimensions, it has provided a model and an inspiration for innumerable local-global results in higher-dimensional geometry, some of which we will prove in Chapter 11.

This chapter begins with some not-so-elementary notions from plane geometry, leading up to a proof of Hopf's rotation angle theorem, which expresses the intuitive idea that the tangent vector of a simple closed curve, or more generally of a "curved polygon," makes a net rotation through an angle of exactly 2π as one traverses the curve counterclockwise. Then we investigate curved polygons on Riemannian 2-manifolds, leading to a far-reaching generalization of the rotation angle theorem called the Gauss–Bonnet formula, which expresses the relationship among the exterior angles, the geodesic curvature of the boundary, and the Gaussian curvature in the interior of a curved polygon. Finally, we use the Gauss–Bonnet formula to prove the global statement of the Gauss–Bonnet theorem.

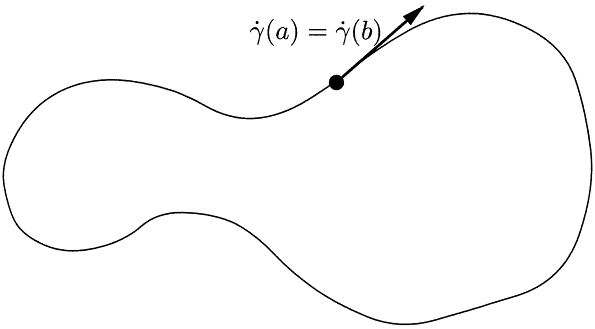


FIGURE 9.1. A closed curve with $\dot\gamma(a) = \dot\gamma(b)$.

Some Plane Geometry

Look back for a moment at the three local-global theorems about plane geometry stated in Chapter 1: the angle-sum theorem, the circumference theorem, and the total curvature theorem. When looked at correctly, these three theorems all turn out to be manifestations of the same phenomenon: as one traverses a simple closed plane curve in the counterclockwise direction, the tangent vector makes a net rotation through an angle of exactly 2π. Our task in the first part of this chapter is to make these notions precise.

Throughout this section, $\gamma\colon [a,b] \to \mathbf{R}^2$ is a unit speed admissible curve in the plane. We say γ is *simple* if it is injective on $[a,b)$, and *closed* if $\gamma(b) = \gamma(a)$.

If γ is smooth, we can define the *tangent angle* $\theta(t)$ as the unique continuous map $\theta\colon [a,b] \to \mathbf{R}$ such that $\dot\gamma(t) = (\cos\theta(t), \sin\theta(t))$ for all $t \in [a,b]$, and such that $\theta(a) \in (-\pi,\pi]$. That such a continuous choice of angle exists follows from the theory of covering spaces: since γ is unit speed, and the tangent space to $\mathbf{R}^2$ is naturally identified with $\mathbf{R}^2$ itself, we can think of $\dot\gamma$ as a map from $[a,b]$ to $\mathbf{S}^1$. By the path-lifting property of covering maps [Mas67, Lemma V.3.1], this map lifts to the universal covering $\pi\colon \mathbf{R} \to \mathbf{S}^1$ given by $\pi(\theta) = (\cos\theta, \sin\theta)$. Our tangent angle function θ is the unique continuous lift with the additional property that $\theta(a) \in (-\pi,\pi]$. Because $\dot\gamma\colon [a,b] \to \mathbf{S}^1$ is a smooth map, and the covering map π is a local diffeomorphism, it follows that θ is actually smooth.

If γ is a unit speed regular closed curve such that $\dot\gamma(a) = \dot\gamma(b)$ (Figure 9.1), we define the *rotation angle* of γ to be $\operatorname{Rot}(\gamma) := \theta(b) - \theta(a)$, where θ is the tangent angle function defined above. Clearly $\operatorname{Rot}(\gamma)$ is an integral multiple of 2π, since $\theta(a)$ and $\theta(b)$ both represent the angle from the x-axis to $\dot\gamma(a)$. (Note that our choice of normalization $\theta(a) \in (-\pi,\pi]$ is immaterial here; we just chose it so that θ would be uniquely defined.)

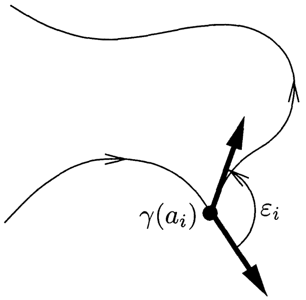


FIGURE 9.2. An exterior angle.

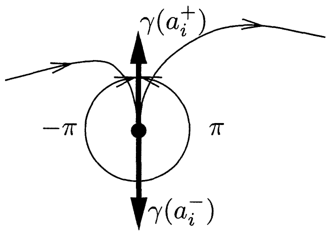


FIGURE 9.3. A cusp.

We would also like to extend the definition of the rotation angle to certain piecewise smooth closed curves. For this purpose, we have to take into account the "jumps" in the tangent angle at corners. To do so, recall that γ has left-hand and right-hand tangent vectors at $t = a_i$, denoted $\dot\gamma(a_i^-)$ and $\dot\gamma(a_i^+)$, respectively. Define the *exterior angle* at a_i to be the oriented angle ε_i from $\dot\gamma(a_i^-)$ to $\dot\gamma(a_i^+)$, chosen to be in the interval $[-\pi, \pi]$, with a positive sign if $(\dot\gamma(a_i^-), \dot\gamma(a_i^+))$ is an oriented basis for $\mathbf{R}^2$, and a negative sign otherwise (Figure 9.2). (If $\dot\gamma(a_i^-) = -\dot\gamma(a_i^+)$, γ has a "cusp" and there is no unambiguous way to choose between π and $-\pi$ (Figure 9.3); for now we leave it unspecified.) If γ is closed, define the exterior angle at $\gamma(a) = \gamma(b)$ to be the angle from $\dot\gamma(b)$ to $\dot\gamma(a)$, chosen in the interval $[-\pi, \pi]$.

The curves we wish to consider are of the following type: A *curved polygon* in the plane is a simple, closed, piecewise smooth, unit speed curve segment, none of whose exterior angles is equal to $\pm\pi$, that is the boundary of a bounded open set $\Omega \subset \mathbf{R}^2$. If $a = a_0 < \cdots < a_k = b$ is a subdivision of $[a, b]$ such that γ is smooth on $[a_{i-1}, a_i]$, the points $\gamma(a_i)$ are called the *vertices* of γ, and the curve segments $\gamma|_{[a_{i-1}, a_i]}$ are called its *edges* or *sides.*

If γ is parametrized so that at points where γ is smooth, $\dot\gamma$ is consistent with the induced orientation on $\gamma = \partial\Omega$ in the sense of Stokes's theorem, we say γ is *positively oriented* (Figure 9.4). Intuitively, this just means that γ is parametrized in the counterclockwise direction, or that Ω is always to the left of γ.

Suppose γ is a curved polygon. We define the *tangent angle* $\theta\colon [a, b] \to \mathbf{R}$ as follows (Figures 9.5 and 9.6): Beginning with $\theta(a) \in (-\pi, \pi]$, we define $\theta(t)$ for $t \in [a, a_1)$ to be the unique continuous choice of angle from the x-axis to $\dot\gamma(t)$ as above. At the first vertex $\gamma(a_1)$, let

$$\theta(a_1) = \lim_{t \nearrow a_1} \theta(t) + \varepsilon_1.$$

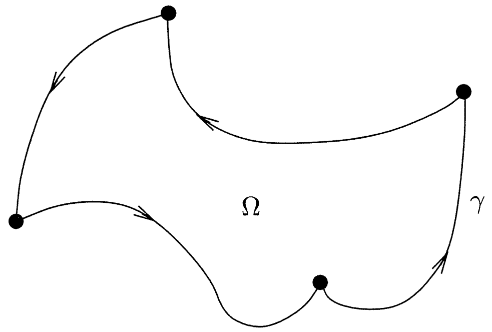


FIGURE 9.4. A positively oriented curved polygon.

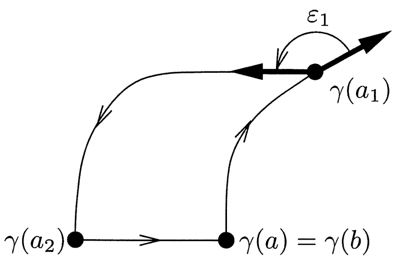


FIGURE 9.5. Defining the tangent angle at a vertex.

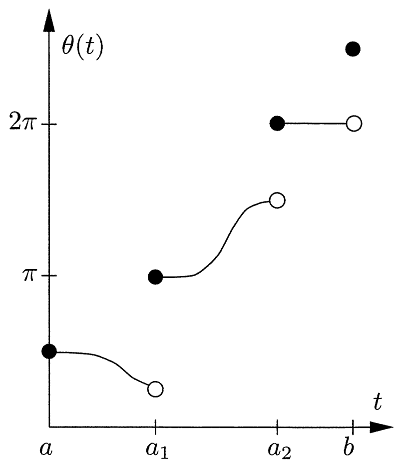


FIGURE 9.6. The tangent angle function for the curve in Fig. 9.5.

(See Figure 9.5.) Then extend θ continuously on $[a_1, a_2)$, and continue by induction, until finally

$$\theta(b) = \lim_{t \nearrow b} \theta(t) + \varepsilon_k,$$

where ε_k is the exterior angle at $\gamma(b)$. We define the rotation angle of γ to be $\mathrm{Rot}(\gamma) := \theta(b) - \theta(a)$. $\mathrm{Rot}(\gamma)$ is again an integral multiple of 2π, because the definition ensures that $\theta(b)$ and $\theta(a)$ are both representations of the angle from the x-axis to $\dot{\gamma}(a)$.

The following theorem is due to Heinz Hopf [Hop35] (for a more accessible version of the proof, see [Hop83, formula (7.1)]). In the literature, it is sometimes referred to by the German name given to it by Hopf, the *Umlaufsatz.*

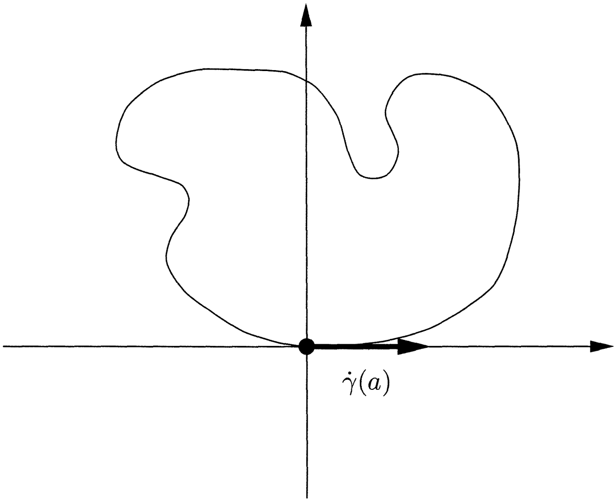


FIGURE 9.7. The curve γ after changing the parameter interval and translating $\gamma(a)$ to the origin.

Theorem 9.1. (Rotation Angle Theorem) *If γ is a positively oriented curved polygon in the plane, the rotation angle of γ is exactly 2π.*

Proof. Suppose first that all the exterior angles are zero. This means, in particular, that $\dot\gamma$ is continuous and $\dot\gamma(a) = \dot\gamma(b)$. Since γ is closed, we can extend it to a continuous map from $\mathbf{R}$ to $\mathbf{R}^2$ by requiring it to be periodic of period $b-a$. Our hypothesis that $\dot\gamma(a) = \dot\gamma(b)$ guarantees that the extended map still has continuous first derivatives.

$\operatorname{Rot}(\gamma)$ is clearly unchanged if we consider γ as being defined on any interval $[\tilde a, \tilde b]$ of length $b - a$ (this just changes the point at which we start). Let's choose our parameter interval $[\tilde a, \tilde b]$ such that the y-coordinate of γ achieves its minimum at $t = \tilde a$; for convenience, we relabel the new interval as $[a, b]$. Moreover, by a translation we may as well assume that $\gamma(a)$ is the origin. Then the image of γ remains in the upper half-plane, and $\dot\gamma(a) = \dot\gamma(b) = \partial/\partial x$ (Figure 9.7).

Since $\dot\gamma$ is continuous, so is the tangent angle function $\theta\colon [a,b] \to \mathbf{R}$. We will extend this function to a continuous *secant angle function* $\varphi(t_1,t_2)$ defined on the triangle $T := \{(t_1,t_2) : a \le t_1 \le t_2 \le b\}$ (Figure 9.8), representing the angle between the x-axis and the vector from $\gamma(t_1)$ to

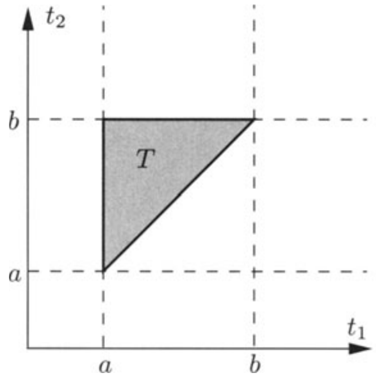


FIGURE 9.8. The domain of the secant angle function.

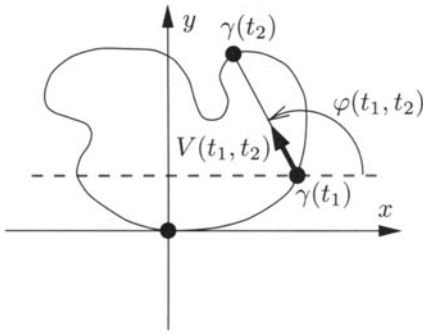


FIGURE 9.9. The secant angle.

$\gamma(t_2)$. To be precise, first define a map $V\colon T \to \mathbf{S}^1$ by

$$V(t_1,t_2) = \begin{cases} \dfrac{\gamma(t_2)-\gamma(t_1)}{|\gamma(t_2)-\gamma(t_1)|}, & t_1 < t_2 \text{ and } (t_1,t_2) \neq (a,b); \\ \dot\gamma(t_1), & t_1 = t_2; \\ -\dot\gamma(a), & (t_1,t_2) = (a,b). \end{cases}$$

V is continuous along the line $t_1 = t_2$, because

$$\begin{aligned} \lim_{(t_1,t_2)\to(t,t)} \frac{\gamma(t_2)-\gamma(t_1)}{|\gamma(t_2)-\gamma(t_1)|} &= \lim_{(t_1,t_2)\to(t,t)} \frac{\gamma(t_2)-\gamma(t_1)}{t_2-t_1} \Big/ \left|\frac{\gamma(t_2)-\gamma(t_1)}{t_2-t_1}\right| \\ &= \dot\gamma(t)/|\dot\gamma(t)| \\ &= V(t,t). \end{aligned}$$

A similar argument shows that V is continuous at (a,b), using the fact that $\dot\gamma(a) = \dot\gamma(b)$. Since T is simply connected, the theory of covering spaces (cf. [Mas67, Theorem V.5.1]) guarantees that $V\colon T \to \mathbf{S}^1$ has a continuous lift $\varphi\colon T \to \mathbf{R}$, which is unique if we require $\varphi(a,a) = 0$ (Figure 9.9). This is our secant angle function.

We can write $\operatorname{Rot}(\gamma) = \theta(b) - \theta(a) = \varphi(b,b) - \varphi(a,a) = \varphi(b,b)$. Observe that, along the side of T where $t_1 = a$ and $t_2 \in [a,b]$, the vector V has its tail at the origin and its head in the upper half-plane. Since we stipulate that $\varphi(a,a) = 0$, we must have $\varphi(a,t_2) \in [0,\pi]$ on this segment. By continuity, therefore, $\varphi(a,b) = \pi$ (since $\varphi(a,b)$ represents the angle of $-\dot\gamma(a) = -\partial/\partial x$). Similarly, on the side where $t_2 = b$, V has its head at the origin and its tail in the upper half-plane, so $\varphi(t_1,b) \in [\pi, 2\pi]$. Therefore, since $\varphi(b,b)$ represents the angle of $\dot\gamma(b) = \partial/\partial x$, we must have $\varphi(b,b) = 2\pi$. This completes the proof for the case where $\dot\gamma$ is continuous.

Now suppose γ has vertices. It suffices to show there is a curve with a continuous tangent vector that has the same rotation angle as γ. We will

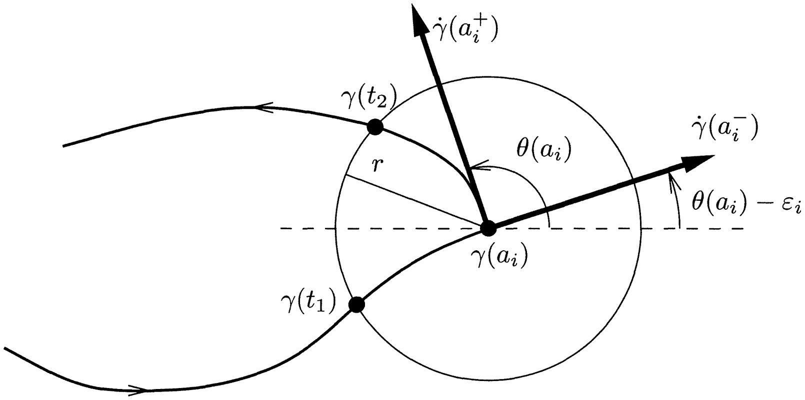


FIGURE 9.10. Isolating the change in the tangent angle at a vertex.

construct such a curve by "rounding the corners" of γ. It will simplify the proof somewhat if we choose the parameter interval $[a, b]$ so that $\gamma(a) = \gamma(b)$ is not a vertex.

Let $\gamma(a_i)$ be any vertex, and ε_i its exterior angle. Let α be a small positive number depending on ε_i; we will describe how to choose it later. Recall that our definition of $\theta(t)$ guarantees that θ is continuous from the right, and $\lim_{t \nearrow a_i} \theta(t) = \theta(a_i) - \varepsilon_i$. Therefore, we can choose δ small enough that $|\theta(t) - (\theta(a_i) - \varepsilon_i)| < \alpha$ when $t \in (a_i - \delta, a_i)$, and $|\theta(t) - \theta(a_i)| < \alpha$ when $t \in (a_i, a_i + \delta)$.

The image under γ of $[a, b] - (a_i - \delta, a_i + \delta)$ is a compact set disjoint from $\gamma(a_i)$, so we can choose r small enough that γ does not enter $\overline{B}_r(\gamma(a_i))$ except when $t \in (a_i - \delta, a_i + \delta)$. Let $t_1 \in (a_i - \delta, a_i + \delta)$ denote the time when γ enters $\overline{B}_r(\gamma(a_i))$, and t_2 the time when it leaves (Figure 9.10). By our choice of δ, the total change in $\theta(t)$ is not more than α when $t \in [t_1, a_i)$, and again not more than α when $t \in (a_i, t_2]$. Therefore, if α is small enough, the total change $\Delta\theta$ in $\theta(t)$ during the time interval $[t_1, t_2]$ is between $\varepsilon_i - 2\alpha$ and $\varepsilon_i + 2\alpha$. If we choose $\alpha < \frac{1}{2}(\pi - |\varepsilon_i|)$, it satisfies $-\pi < \Delta\theta < \pi$.

Now we simply replace the portion of γ from time t_1 to time t_2 with a smooth curve segment σ that is tangent to γ at $\gamma(t_1)$ and $\gamma(t_2)$, and whose tangent angle increases or decreases monotonically from $\theta(t_1)$ to $\theta(t_2)$; an arc of a hyperbola will do (Figure 9.11). Since the change in tangent angle of σ is between $-\pi$ and π and represents the angle between $\dot\gamma(t_1)$ and $\dot\gamma(t_2)$, it must be exactly $\Delta\theta$. (The length of σ may not be the same as that of the portion of γ being replaced, but we can simply reparametrize the new curve by arc length.) Repeating this process for each vertex, we obtain a new curve with a continuous tangent vector field whose rotation angle is the same as that of γ, thus proving the theorem. □

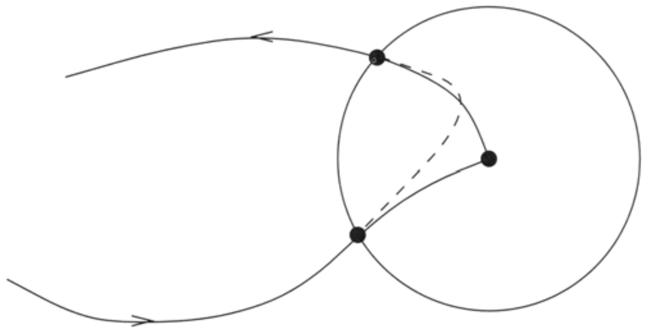


FIGURE 9.11. Rounding a corner.

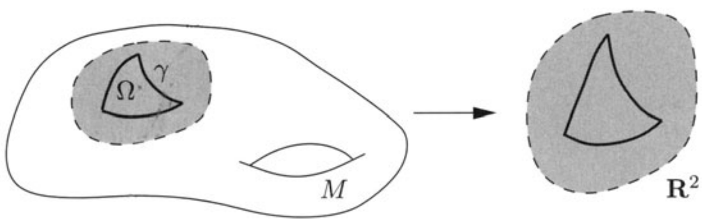


FIGURE 9.12. A curved polygon on a surface.

From the rotation angle theorem, it is not hard to deduce the three local-global theorems mentioned at the beginning of the chapter as corollaries. (The angle-sum theorem is trivial; for the total curvature theorem, the trick is to show that $\dot{\theta}(t)$ is equal to the signed curvature of γ; the circumference theorem follows from the total curvature theorem as mentioned in Chapter 1.) However, instead of proving them directly, we will prove a general formula, called the Gauss–Bonnet formula, from which these results and more follow easily. You will easily see how the statement and proof of Theorem 9.3 below can be simplified in case the metric is Euclidean.

The Gauss–Bonnet Formula

We now direct our attention to the case of an oriented Riemannian 2-manifold (M, g). In this setting, a unit speed curve $\gamma\colon [a,b] \to M$ is called a curved polygon if γ is the boundary of an open set Ω with compact closure, and there is a coordinate chart containing γ and Ω under whose image γ is a curved polygon in the plane (Figure 9.12). Using the coordinates to transfer γ, Ω, and g to the plane, we may as well assume that g is a metric on some open subset $\mathcal{U} \subset \mathbf{R}^2$, and γ is a curved polygon in $\mathcal{U}$.

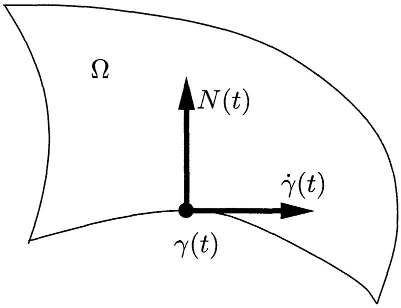


FIGURE 9.13. $N(t)$ is the inward-pointing normal.

For a curved polygon γ in M, our previous definitions go through almost unchanged. We say γ is positively oriented if it is parametrized in the direction of its Stokes orientation as the boundary of Ω. We define the exterior angle ε_i at a vertex $\gamma(a_i)$ as the oriented angle from $\dot\gamma(a_i^-)$ to $\dot\gamma(a_i^+)$ with respect to the g-inner product and the given orientation of M, so it satisfies $\cos\varepsilon_i = \langle\dot\gamma(a_i^+), \dot\gamma(a_i^-)\rangle$. Having chosen coordinates, we define the tangent angle $\theta\colon [a,b] \to \mathbf{R}$ on segments where $\dot\gamma$ is continuous as the unique continuous choice of angle from $\partial/\partial x$ to $\dot\gamma$, measured with respect to g, with jumps at vertices as before. The rotation angle is $\operatorname{Rot}(\gamma) = \theta(b) - \theta(a)$. Because of the role played by $\partial/\partial x$ in the definition, it is not clear yet that the rotation angle has any coordinate-invariant meaning; however, we do have the following easy consequence of the rotation angle theorem.

Lemma 9.2. *If γ is a positively oriented curved polygon in M, the rotation angle of γ is 2π.*

Proof. If we use the given coordinate chart to consider γ as a curved polygon in the plane, we can compute its tangent angle function either with respect to g or with respect to the Euclidean metric $\bar g$. In either case, $\operatorname{Rot}(\gamma)$ is an integral multiple of 2π because $\theta(a)$ and $\theta(b)$ both represent the same angle. Now for $0 \le s \le 1$, let $g_s = sg + (1-s)\bar g$. By the same reasoning, the rotation angle $\operatorname{Rot}_{g_s}(\gamma)$ with respect to g_s is also a multiple of 2π. The function $f(s) = (1/2\pi)\operatorname{Rot}_{g_s}(\gamma)$ is therefore integer-valued, and is easily seen to be continuous in s, so it must be constant. □

There is a unique unit normal vector field along the smooth portions of γ such that $(\dot\gamma(t), N(t))$ is an oriented orthonormal basis for $T_{\gamma(t)}M$ for each t. If γ is positively oriented as the boundary of Ω, this is equivalent to N being the inward-pointing normal to $\partial\Omega$ (Figure 9.13). We define the *signed curvature* $\kappa_N(t)$ at smooth points of γ by

$$\kappa_N(t) = \langle D_t\dot\gamma(t), N(t)\rangle\,.$$

By differentiating $|\dot\gamma(t)|^2 \equiv 1$, we see that $D_t\dot\gamma(t)$ is orthogonal to $\dot\gamma(t)$, and therefore we can write $D_t\dot\gamma(t) = \kappa_N(t)N(t)$, and the (unsigned) curvature of γ is $\kappa(t) = |\kappa_N(t)|$. The sign of $\kappa_N(t)$ is positive if γ is curving toward Ω, and negative if it is curving away.

Theorem 9.3. (The Gauss–Bonnet Formula) *Suppose γ is a curved polygon on an oriented Riemannian 2-manifold (M,g), and γ is positively oriented as the boundary of an open set Ω with compact closure. Then*

$$\int_\Omega K\,dA + \int_\gamma \kappa_N\,ds + \sum_i \varepsilon_i = 2\pi, \tag{9.1}$$

where K is the Gaussian curvature of g and dA is its Riemannian volume element.

Proof. Let $a = a_0 < \cdots < a_k = b$ be a subdivision of $[a,b]$ into segments on which γ is smooth. Using the rotation angle theorem and the fundamental theorem of calculus, we can write

$$2\pi = \sum_{i=1}^k \varepsilon_i + \sum_{i=1}^k \int_{a_{i-1}}^{a_i} \dot\theta(t)\,dt. \tag{9.2}$$

To prove (9.1), we need to derive a relationship among $\dot\theta$, κ_N, and K.

We begin by constructing a specially adapted orthonormal frame. Let (x,y) be oriented coordinates on an open set $\mathcal{U}$ containing γ and Ω. The Gram–Schmidt algorithm applied to the frame $(\partial/\partial x, \partial/\partial y)$ yields an oriented orthonormal frame (E_1,E_2) such that E_1 is a positive multiple of $\partial/\partial x$. Then, because $\theta(t)$ represents the g-angle between E_1 and $\dot\gamma(t)$, it is easy to see that the following hold at smooth points of γ:

$$\begin{aligned}\dot\gamma(t) &= \cos\theta(t)E_1 + \sin\theta(t)E_2;\\ N(t) &= -\sin\theta(t)E_1 + \cos\theta(t)E_2.\end{aligned}$$

Differentiating $\dot\gamma$ (and omitting the t dependence from the notation for simplicity), we get

$$\begin{aligned}D_t\dot\gamma &= -\dot\theta(\sin\theta)E_1 + (\cos\theta)\nabla_{\dot\gamma}E_1 + \dot\theta(\cos\theta)E_2 + (\sin\theta)\nabla_{\dot\gamma}E_2\\ &= \dot\theta N + (\cos\theta)\nabla_{\dot\gamma}E_1 + (\sin\theta)\nabla_{\dot\gamma}E_2.\end{aligned} \tag{9.3}$$

Next we analyze the covariant derivatives of E_1 and E_2. Because (E_1,E_2) is an orthonormal frame, for any vector X we have

$$\begin{aligned}0 &= \nabla_X|E_1|^2 = 2\langle\nabla_X E_1, E_1\rangle\\ 0 &= \nabla_X|E_2|^2 = 2\langle\nabla_X E_2, E_2\rangle\\ 0 &= \nabla_X\langle E_1,E_2\rangle = \langle\nabla_X E_1, E_2\rangle + \langle E_1, \nabla_X E_2\rangle.\end{aligned}$$

The first two equations show that $\nabla_X E_1$ is a multiple of E_2 and $\nabla_X E_2$ is a multiple of E_1. Define a 1-form ω by

$$\omega(X) := \langle E_1, \nabla_X E_2 \rangle = -\langle \nabla_X E_1, E_2 \rangle.$$

It follows that the covariant derivatives of the basis elements are given by

$$\begin{aligned} \nabla_X E_1 &= -\omega(X) E_2; \\ \nabla_X E_2 &= \omega(X) E_1. \end{aligned} \tag{9.4}$$

Thus the 1-form ω completely determines the connection in $\mathcal{U}$. (In fact, when the connection is expressed in terms of the local frame $\{E_i\}$ as in Problem 4-5, this computation shows that the connection 1-forms are just ${\omega_2}^1 = -{\omega_1}^2 = \omega$, ${\omega_1}^1 = {\omega_2}^2 = 0$.)

Using (9.3) and (9.4), we can compute

$$\begin{aligned} \kappa_N &= \langle D_t \dot\gamma, N \rangle \\ &= \langle \dot\theta N, N \rangle + \cos\theta \langle \nabla_{\dot\gamma} E_1, N \rangle + \sin\theta \langle \nabla_{\dot\gamma} E_2, N \rangle \\ &= \dot\theta - \cos\theta \langle \omega(\dot\gamma) E_2, N \rangle + \sin\theta \langle \omega(\dot\gamma) E_1, N \rangle \\ &= \dot\theta - \cos^2\theta\, \omega(\dot\gamma) - \sin^2\theta\, \omega(\dot\gamma) \\ &= \dot\theta - \omega(\dot\gamma). \end{aligned}$$

Therefore, (9.2) becomes

$$\begin{aligned} 2\pi &= \sum_{i=1}^k \varepsilon_i + \sum_{i=1}^k \int_{a_{i-1}}^{a_i} \kappa_N(t)\, dt + \sum_{i=1}^k \int_{a_{i-1}}^{a_i} \omega(\dot\gamma(t))\, dt \\ &= \sum_{i=1}^k \varepsilon_i + \int_\gamma \kappa_N\, ds + \int_\gamma \omega. \end{aligned}$$

The theorem will therefore be proved if we can show that

$$\int_\gamma \omega = \int_\Omega K\, dA. \tag{9.5}$$

If γ were a smooth closed curve, Stokes's theorem would imply that the left-hand side of (9.5) is equal to $\int_\Omega d\omega$. In fact, this is true anyway: by a construction similar to that used in the proof of the rotation angle theorem, we can approximate γ uniformly by a sequence of smooth curves γ_j whose lengths approach that of γ, and that are boundaries of domains Ω_j such that the area between Ω_j and Ω approaches zero. Applying Stokes's theorem on Ω_j and taking the limit as $j \to \infty$, we conclude that

$$2\pi = \sum_{i=1}^k \varepsilon_i + \int_\gamma \kappa_N\, ds + \int_\Omega d\omega.$$

The last step of the proof is to show that $d\omega = K\,dA$. This follows from the general formula relating the curvature tensor and the connection 1-forms given in Problem 7-2; but in the case of two dimensions we can give an easy direct proof. Since (E_1, E_2) is an oriented orthonormal frame, it follows by definition of the Riemannian volume element that $dA(E_1, E_2) = 1$. Using (9.4), we compute

$$\begin{aligned}
K\,dA(E_1,E_2) &= K = Rm(E_1,E_2,E_2,E_1)\\
&= \langle \nabla_{E_1}\nabla_{E_2}E_2 - \nabla_{E_2}\nabla_{E_1}E_2 - \nabla_{[E_1,E_2]}E_2, E_1\rangle\\
&= \langle \nabla_{E_1}(\omega(E_2)E_1) - \nabla_{E_2}(\omega(E_1)E_1) - \omega[E_1,E_2]E_1, E_1\rangle\\
&= \langle E_1(\omega(E_2))E_1 + \omega(E_2)\nabla_{E_1}E_1 - E_2(\omega(E_1))E_1\\
&\qquad - \omega(E_1)\nabla_{E_2}E_1 - \omega[E_1,E_2]E_1, E_1\rangle\\
&= E_1(\omega(E_2)) - E_2(\omega(E_1)) - \omega[E_1,E_2]\\
&= d\omega(E_1,E_2).
\end{aligned}$$

This completes the proof. □

The three local-global theorems of plane geometry stated in Chapter 1 follow from the Gauss–Bonnet formula as easy corollaries. Their proofs are left to the reader.

Corollary 9.4. (Angle-Sum Theorem) *The sum of the interior angles of a Euclidean triangle is π.*

Corollary 9.5. (Circumference Theorem) *The circumference of a Euclidean circle of radius R is $2\pi R$.*

Corollary 9.6. (Total Curvature Theorem) *If $\gamma\colon [a,b] \to \mathbf{R}^2$ is a unit speed simple closed curve such that $\dot\gamma(a) = \dot\gamma(b)$, and N is the inward-pointing normal, then*

$$\int_a^b \kappa_N(t)\,dt = 2\pi.$$

Exercise 9.1. Prove the three corollaries above.

The Gauss–Bonnet Theorem

It is now a relatively easy matter to "globalize" the Gauss–Bonnet formula to obtain the Gauss–Bonnet theorem. The link between the local and global results is provided by triangulations, so we begin by discussing this construction borrowed from algebraic topology. Most of the topological ideas touched upon in this section can be found treated in detail in either [Sie92] or [Mas67].

If M is a smooth, compact 2-manifold, a *smooth triangulation* of M is a finite collection of curved triangles (i.e., three-sided curved polygons),

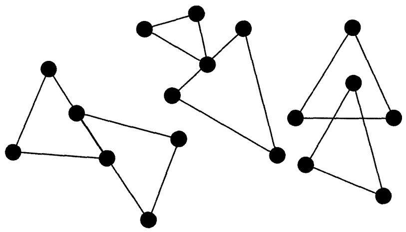


FIGURE 9.14. Illegal intersections of triangles in a triangulation.

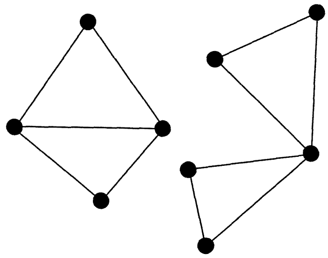


FIGURE 9.15. Valid intersections.

such that the union of the closed regions $\overline{\Omega}_i$ bounded by the triangles is M, and the intersection of any pair (if not empty) is either a single vertex of each or a single edge of each (Figures 9.14 and 9.15). Every smooth, compact surface possesses a smooth triangulation. In fact, it was proved by Tibor Radó [Rad25] in 1925 that every compact *topological* 2-manifold possesses a triangulation (without the assumption of smoothness of the edges, of course). There is a proof for the smooth case that is not terribly hard, outlined in Problem 9-5.

If M is a triangulated 2-manifold, the *Euler characteristic* of M with respect to the given triangulation is defined to be

$$\chi(M) := N_v - N_e + N_f,$$

where N_v is the number of vertices in the triangulation, N_e is the number of edges, and N_f is the number of faces (the Ω_is). It is an important result of algebraic topology that the Euler characteristic is in fact a topological invariant, and is independent of the choice of triangulation (see [Sie92, Theorem 13.3.1]).

Theorem 9.7. (The Gauss–Bonnet Theorem) *If M is a triangulated, compact, oriented, Riemannian 2-manifold, then*

$$\int_M K\, dA = 2\pi\chi(M).$$

Proof. Let $\{\Omega_i : i = 1, \dots, N_f\}$ denote the faces of the triangulation, and for each i let $\{\gamma_{ij} : j = 1, 2, 3\}$ be the edges of Ω_i and $\{\theta_{ij} : j = 1, 2, 3\}$ its interior angles. Since each exterior angle is π minus the corresponding interior angle, applying the Gauss–Bonnet formula to each triangle and summing over i gives

$$\sum_{i=1}^{N_f} \int_{\Omega_i} K\, dA + \sum_{i=1}^{N_f} \sum_{j=1}^{3} \int_{\gamma_{ij}} \kappa_N\, ds + \sum_{i=1}^{N_f} \sum_{j=1}^{3} (\pi - \theta_{ij}) = \sum_{i=1}^{N_f} 2\pi. \tag{9.6}$$

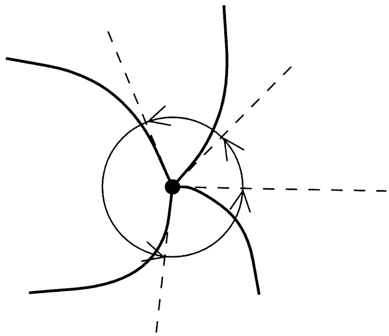


FIGURE 9.16. Interior angles at a vertex add up to 2π.

Note that each edge integral appears exactly twice in the above sum, with opposite orientations, so the integrals of κ_N all cancel out. Thus (9.6) becomes

$$\int_M K\,dA + 3\pi N_f - \sum_{i=1}^{N_f}\sum_{j=1}^{3}\theta_{ij} = 2\pi N_f. \tag{9.7}$$

Note also that each interior angle θ_{ij} appears exactly once. At each vertex, the angles that touch that vertex must add up to 2π (Figure 9.16); thus the angle sum can be rearranged to give exactly $2\pi N_v$. Equation (9.7) thus can be written

$$\int_M K\,dA = 2\pi N_v - \pi N_f. \tag{9.8}$$

Finally, since each edge appears in exactly two triangles, and each triangle has exactly three edges, the total number of edges *counted with multiplicity* is $2N_e = 3N_f$, where we count each edge once for each triangle in which it appears. This means that $N_f = 2N_e - 2N_f$, so (9.8) finally becomes

$$\int_M K\,dA = 2\pi N_v - 2\pi N_e + 2\pi N_f = 2\pi\chi(M).$$

□

The significance of this theorem cannot be overstated. Together with the classification theorem for compact surfaces, it gives us a very complete picture of the possible Gaussian curvatures for metrics on compact surfaces. The classification theorem (see, for example, [Sie92, Theorem 13.2.5] or [Mas67, Theorem I.5.1]) says that every compact, orientable 2-manifold is homeomorphic to a sphere or the connected sum of g tori, and every

nonorientable one is homeomorphic to the connected sum of g copies of the projective plane $\mathbf{P}^2$; the number g is called the *genus* of the surface. (The sphere is said to have genus zero.) By constructing simple triangulations, it is easy to check that the Euler characteristic of an orientable surface of genus g is $2-2g$, and that of a nonorientable one is $2-g$.

Corollary 9.8. *Let M be a compact Riemannian 2-manifold and K its Gaussian curvature.*

(*a*) *If M is homeomorphic to the sphere or the projective plane, then $K>0$ somewhere.*

(*b*) *If M is homeomorphic to the torus or the Klein bottle, then either $K\equiv 0$ or K takes on both positive and negative values.*

(*c*) *If M is any other compact surface, then $K<0$ somewhere.*

Proof. If M is orientable, the result follows immediately from the Gauss–Bonnet theorem, because a function whose integral is positive, negative, or zero must satisfy the claimed sign condition. If M is nonorientable, the result follows by applying the Gauss–Bonnet theorem to the orientable double cover $\pi\colon \widetilde{M}\to M$ with the lifted metric $\tilde{g}=\pi^*g$, using the fact that $\widetilde{M}$ is the sphere if $M=\mathbf{P}^2$, the torus if M is the Klein bottle (which is homeomorphic to the connected sum of two copies of $\mathbf{P}^2$), and otherwise has $\chi(\widetilde{M})<0$. □

This corollary has a remarkable converse, proved in the mid-1970s by Jerry Kazdan and Frank Warner: *If K is any smooth function on a compact 2-manifold M satisfying the necessary sign condition of Corollary 9.8, then there exists a Riemannian metric on M for which K is the Gaussian curvature.* The proof is a deep application of the theory of nonlinear partial differential equations. (See [Kaz85] for a nice expository account.)

In Corollary 9.8 we assumed we knew the topology of M and drew conclusions about the possible curvatures it could support. In the following corollary we reverse our point of view, and use assumptions about the curvature to draw conclusions about the manifold.

Corollary 9.9. *Let M be a compact Riemannian 2-manifold and K its Gaussian curvature.*

(*a*) *If $K>0$, then M is homeomorphic to the sphere or projective plane, and $\pi_1(M)$ is finite.*

(*b*) *If $K\le 0$, then $\pi_1(M)$ is infinite, and M has genus at least 1.*

Exercise 9.2. Prove Corollary 9.9.

Much of the effort in contemporary Riemannian geometry is aimed at generalizing the Gauss–Bonnet theorem and its topological consequences to

higher dimensions. As we will see in the next chapter, most of the interesting results have required the development of different methods. However, there is one rather direct generalization of the Gauss–Bonnet theorem that deserves mention: the Chern–Gauss–Bonnet theorem. This was proved by Hopf in 1925 for an n-manifold embedded in $\mathbf{R}^{n+1}$ with the induced metric, and in 1944 by Chern for abstract Riemannian manifolds (see [Spi79, volume 5] for a complete discussion with references). The theorem asserts that on any oriented vector space there exists a basis-independent function

$$\mathcal{P}\colon \{\text{4-tensors with the symmetries of } Rm\} \to \mathbf{R},$$

called the *Pfaffian,* such that for any oriented compact even-dimensional Riemannian n-manifold M,

$$\int_M \mathcal{P}(Rm)\, dV = \frac{1}{2}\operatorname{Vol}(\mathbf{S}^n)\chi(M).$$

(Here $\chi(M)$ is again the Euler characteristic of M, which can be defined analogously to that of a surface and is a topological invariant.)

In a certain sense, this might be considered a very satisfactory generalization of Gauss–Bonnet. The only problem with this result is that the relationship between the Pfaffian and sectional curvatures is obscure in higher dimensions, so no one seems to have any idea how to interpret the theorem geometrically! For example, it is not even known whether the assumption that M has strictly positive sectional curvatures implies that $\chi(M) > 0$.

Problems

9-1. Let $M \subset \mathbf{R}^3$ be a compact, orientable, embedded 2-manifold with the induced metric.

(a) Show that M cannot have $K \leq 0$ everywhere. [Hint: Look at a point where the distance from the origin takes a maximum.]

(b) Show that M cannot have $K \geq 0$ everywhere unless $\chi(M) > 0$.

9-2. Let (M, g) be a Riemannian 2-manifold. A curved polygon on M whose sides are geodesic segments is called a *geodesic polygon.* If g has everywhere nonpositive Gaussian curvature, prove that there are no geodesic polygons with exactly 0, 1, or 2 vertices. Give examples of all three if the curvature hypothesis is not satisfied.

9-3. A *geodesic triangle* on a Riemannian 2-manifold (M, g) is a three-sided geodesic polygon (Problem 9-2).

(a) If M has constant Gaussian curvature K, show that the sum of the interior angles of a geodesic triangle γ is equal to $\pi + KA$, where A is the area of the region bounded by γ.

(b) Suppose M is either the 2-sphere of radius R or the hyperbolic plane of radius R. Show that similar triangles are congruent. More precisely, if γ_1 and γ_2 are geodesic triangles with equal interior angles, then there exists an isometry of M taking γ_1 to γ_2.

9-4. An *ideal triangle* in the hyperbolic plane $\mathbf{H}^2$ is a region whose boundary consists of three geodesics, any two of which meet at a common point on the boundary of the disk (in the Poincaré disk model). Show that all ideal triangles have the same finite area, and compute it. Be careful to justify any limits.

9-5. This problem outlines a proof that every compact smooth 2-manifold has a smooth triangulation.

(a) Show that it suffices to prove there exist finitely many convex geodesic polygons whose interiors cover M, and each of which lies in a uniformly normal convex geodesic ball. (A geodesic polygon is called convex if it together with its interior is a convex set in the sense of Problem 6-4.)

(b) Using the result of Problem 6-4, show that there exist finitely many points $(v_1, \ldots, v_k)$ and $\varepsilon > 0$ such that the geodesic balls $B_{3\varepsilon}(v_i)$ are convex and uniformly normal, and the balls $B_\varepsilon(v_i)$ cover M.

(c) For each i, show that there is a convex geodesic polygon in $B_{3\varepsilon}(v_i)$ whose interior contains $B_\varepsilon(v_i)$. [Hint: Let the vertices be sufficiently nearby points on the circle of radius 2ε around v_i.]

(d) Prove the result.

9-6. Prove the plane curve classification theorem (Theorem 1.5). [Hint: Any plane curve satisfies the ordinary differential equation $\ddot{\gamma}(t) = \kappa_N(t)N(t)$.]

10
Jacobi Fields

Our goal for the remainder of this book is to generalize to higher dimensions some of the geometric and topological consequences of the Gauss–Bonnet theorem. We need to develop a new approach: instead of using Stokes's theorem and differential forms to relate the curvature to global topology as in the proof of the Gauss–Bonnet theorem, we study how curvature affects the behavior of nearby geodesics. Roughly speaking, positive curvature causes nearby geodesics to converge (Figure 10.1), while negative curvature causes them to spread out (Figure 10.2). In order to draw topological consequences from this fact, we need a quantitative way to measure the effect of curvature on a one-parameter family of geodesics.

We begin by deriving the Jacobi equation, which is an ordinary differential equation satisfied by the variation field of any one-parameter family of geodesics. A vector field satisfying this equation along a geodesic is called a Jacobi field. We then introduce the notion of conjugate points, which are pairs of points along a geodesic where some Jacobi field vanishes. Intuitively, if p and q are conjugate along a geodesic, one expects to find a one-parameter family of geodesics that start at p and end (almost) at q.

After defining conjugate points, we prove a simple but essential fact: the points conjugate to p are exactly the points where $\exp_p$ fails to be a local diffeomorphism. We then derive an expression for the second derivative of the length functional with respect to proper variations of a geodesic, called the "second variation formula." Using this formula, we prove another essential fact about conjugate points: No geodesic is minimizing past its first conjugate point.

In the final chapter, we will derive topological consequences of these facts.

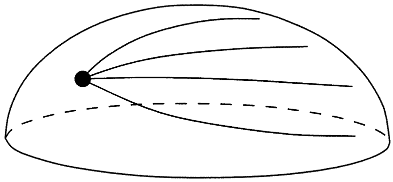


FIGURE 10.1. Positive curvature causes geodesics to converge.

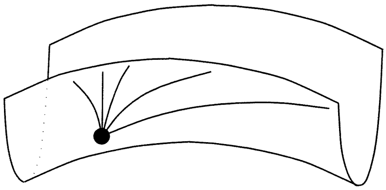


FIGURE 10.2. Negative curvature causes geodesics to spread out.

The Jacobi Equation

In order to study the effect of curvature on nearby geodesics, we focus on variations through geodesics. Suppose therefore that $\gamma\colon [a,b] \to M$ is a geodesic segment, and $\Gamma\colon (-\varepsilon,\varepsilon) \times [a,b] \to M$ is a variation of γ (as defined in Chapter 6). We say Γ is a *variation through geodesics* if each of the main curves $\Gamma_s(t) = \Gamma(s,t)$ is also a geodesic segment. (In particular, this requires that Γ be smooth.) Our first goal is to derive an equation that must be satisfied by the variation field of a variation through geodesics.

Write $T(s,t) = \partial_t\Gamma(s,t)$ and $S(s,t) = \partial_s\Gamma(s,t)$ as in Chapter 6. The geodesic equation tells us that

$$D_tT \equiv 0$$

for all (s,t). We can take the covariant derivative of this equation with respect to s, yielding

$$D_sD_tT \equiv 0.$$

To relate this to the variation field of γ, we need to commute the covariant differentiation operators D_s and D_t. Because these are covariant derivatives acting on a vector field along a curve, we should expect the curvature to be involved. Indeed, we have the following lemma.

Lemma 10.1. *If Γ is any smooth admissible family of curves, and V is a smooth vector field along Γ, then*

$$D_sD_tV - D_tD_sV = R(S,T)V.$$

Proof. This is a local issue, so we can compute in any local coordinates. Writing $V(s,t) = V^i(s,t)\partial_i$, we compute

$$D_tV = \frac{\partial V^i}{\partial t}\partial_i + V^iD_t\partial_i.$$

Therefore,

$$D_sD_tV = \frac{\partial^2 V^i}{\partial s\partial t}\partial_i + \frac{\partial V^i}{\partial t}D_s\partial_i + \frac{\partial V^i}{\partial s}D_t\partial_i + V^iD_sD_t\partial_i.$$

Interchanging D_s and D_t and subtracting, we see that all the terms except the last cancel:

$$D_s D_t V - D_t D_s V = V^i \left(D_s D_t \partial_i - D_t D_s \partial_i \right). \tag{10.1}$$

Now we need to compute the commutator in parentheses. If we write the coordinate functions of Γ as $x^j(s,t)$, then

$$S = \frac{\partial x^k}{\partial s} \partial_k; \qquad T = \frac{\partial x^j}{\partial t} \partial_j.$$

Because ∂_i is extendible,

$$D_t \partial_i = \nabla_T \partial_i = \frac{\partial x^j}{\partial t} \nabla_{\partial_j} \partial_i,$$

and therefore, because $\nabla_{\partial_j} \partial_i$ is also extendible,

$$\begin{aligned} D_s D_t \partial_i &= D_s \left(\frac{\partial x^j}{\partial t} \nabla_{\partial_j} \partial_i \right) \\ &= \frac{\partial^2 x^j}{\partial s \partial t} \nabla_{\partial_j} \partial_i + \frac{\partial x^j}{\partial t} \nabla_S \left(\nabla_{\partial_j} \partial_i \right) \\ &= \frac{\partial^2 x^j}{\partial s \partial t} \nabla_{\partial_j} \partial_i + \frac{\partial x^j}{\partial t} \frac{\partial x^k}{\partial s} \nabla_{\partial_k} \nabla_{\partial_j} \partial_i. \end{aligned}$$

Interchanging $s \leftrightarrow t$ and $j \leftrightarrow k$ and subtracting, we find that the first terms cancel out, and we get

$$\begin{aligned} D_s D_t \partial_i - D_t D_s \partial_i &= \frac{\partial x^j}{\partial t} \frac{\partial x^k}{\partial s} \left(\nabla_{\partial_k} \nabla_{\partial_j} \partial_i - \nabla_{\partial_j} \nabla_{\partial_k} \partial_i \right) \\ &= \frac{\partial x^j}{\partial t} \frac{\partial x^k}{\partial s} R(\partial_k, \partial_j) \partial_i \\ &= R(S,T) \partial_i. \end{aligned}$$

Finally, inserting this into (10.1) yields the result. □

Theorem 10.2. (The Jacobi Equation) *Let γ be a geodesic and V a vector field along γ. If V is the variation field of a variation through geodesics, then V satisfies*

$$D_t^2 V + R(V, \dot\gamma)\dot\gamma = 0. \tag{10.2}$$

Proof. With S and T as before, the preceding lemma implies

$$\begin{aligned} 0 &= D_s D_t T \\ &= D_t D_s T + R(S,T)T \\ &= D_t D_t S + R(S,T)T, \end{aligned}$$

where the last step follows from the symmetry lemma. Evaluating at $s = 0$, where $S(0,t) = V(t)$ and $T(0,t) = \dot\gamma(t)$, we get (10.2). □

Any vector field along a geodesic satisfying the Jacobi equation is called a *Jacobi field.* Because of the following lemma, which is a converse to Theorem 10.2, each Jacobi field tells us how some family of geodesics behaves, at least "infinitesimally" along γ.

Lemma 10.3. *Every Jacobi field along a geodesic γ is the variation field of some variation of γ through geodesics.*

Exercise 10.1. Prove Lemma 10.3. [Hint: Let $\Gamma(s,t) = \exp_{\sigma(s)} tW(s)$ for a suitable curve σ and vector field W along σ.]

Now we reverse our approach: let's forget about variations for a while, and just study Jacobi fields in their own right. As the following lemma shows, the Jacobi equation can be written as a system of second-order linear ordinary differential equations, so it has a unique solution given initial values for V and D_tV at one point.

Proposition 10.4. (Existence and Uniqueness of Jacobi Fields) *Let $\gamma\colon I \to M$ be a geodesic, $a \in I$, and $p = \gamma(a)$. For any pair of vectors $X, Y \in T_pM$, there is a unique Jacobi field J along γ satisfying the initial conditions*

$$J(a) = X; \qquad D_tJ(a) = Y.$$

Proof. Choose an orthonormal basis $\{E_i\}$ for T_pM, and extend it to a parallel orthonormal frame along all of γ. Writing $J(t) = J^i(t)E_i$, we can express the Jacobi equation as

$$\ddot{J}^i + R_{jkl}{}^i J^j \dot{\gamma}^k \dot{\gamma}^l = 0.$$

This is a linear system of second-order ODEs for the n functions J^i. Making the usual substitution $V^i = \dot{J}^i$ converts it to an equivalent first-order linear system for the $2n$ unknowns $\{J^i, V^i\}$. Then Theorem 4.12 guarantees the existence and uniqueness of a solution on the whole interval I with any initial conditions $J^i(a) = X^i$, $V^i(a) = Y^i$. □

Corollary 10.5. *Along any geodesic γ, the set of Jacobi fields is a $2n$-dimensional linear subspace of $\mathcal{T}(\gamma)$.*

Proof. Let $p = \gamma(a)$ be any point on γ, and consider the map from the set of Jacobi fields along γ to $T_pM \oplus T_pM$ by sending J to $(J(a), D_tJ(a))$. The preceding proposition says precisely that this map is bijective. □

There are always two trivial Jacobi fields along any geodesic, which we can write down immediately (see Figure 10.3). Because $D_t\dot{\gamma} = 0$ and $R(\dot{\gamma}, \dot{\gamma})\dot{\gamma} = 0$ by antisymmetry of R, the vector field $J_0(t) = \dot{\gamma}(t)$ satisfies the Jacobi equation with initial conditions

$$J_0(0) = \dot{\gamma}(0); \qquad D_tJ_0(0) = 0.$$

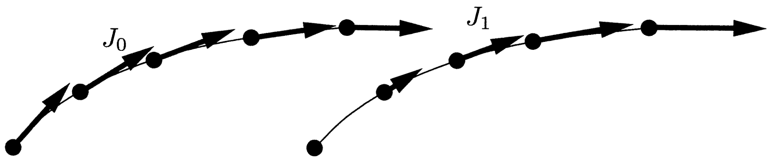


FIGURE 10.3. Trivial Jacobi fields.

Similarly, $J_1(t) = t\dot\gamma(t)$ is a Jacobi field with initial conditions

$$J_1(0) = 0; \qquad D_tJ_1(0) = \dot\gamma(0).$$

It is easy to see that J_0 is the variation field of the variation $\Gamma(s,t) = \gamma(s+t)$, while J_1 is the variation field of $\Gamma(s,t) = \gamma(e^s t)$. Therefore, these two Jacobi fields just reflect the possible reparametrizations of γ, and don't tell us anything about the behavior of geodesics other than γ itself.

To distinguish these trivial cases from more informative ones, we make the following definitions. A *tangential vector field* along a curve γ is a vector field V such that $V(t)$ is a multiple of $\dot\gamma(t)$ for all t, and a *normal vector field* is one such that $V(t) \perp \dot\gamma(t)$ for all t.

Lemma 10.6. *Let $\gamma\colon I \to M$ be a geodesic, and $a \in I$.*

(*a*) *A Jacobi field J along γ is normal if and only if*

$$J(a) \perp \dot\gamma(a) \text{ and } D_tJ(a) \perp \dot\gamma(a). \tag{10.3}$$

(*b*) *Any Jacobi field orthogonal to $\dot\gamma$ at two points is normal.*

Proof. Using compatibility with the metric and the fact that $D_t\dot\gamma \equiv 0$, we compute

$$\begin{aligned}\frac{d^2}{dt^2}\langle J,\dot\gamma\rangle &= \langle D_t^2 J,\dot\gamma\rangle \\ &= -\langle R(J,\dot\gamma)\dot\gamma,\dot\gamma\rangle \\ &= -Rm(J,\dot\gamma,\dot\gamma,\dot\gamma) = 0\end{aligned}$$

by the symmetries of the curvature tensor. Thus, by elementary calculus, $f(t) := \langle J(t),\dot\gamma(t)\rangle$ is a linear function of t. Note that $f(a) = \langle J(a),\dot\gamma(a)\rangle$ and $\dot f(a) = \langle D_tJ(a),\dot\gamma(a)\rangle$. Thus $J(a)$ and $D_tJ(a)$ are orthogonal to $\dot\gamma(a)$ if and only if f and its first derivative vanish at a, which happens if and only if $f \equiv 0$. Similarly, if J is orthogonal to $\dot\gamma$ at two points, then f vanishes at two points and is therefore identically zero. □

As a consequence of this lemma, it is easy to check that the space of normal Jacobi fields is a $(2n-2)$-dimensional subspace of $\mathcal{T}(\gamma)$, and the space of tangential ones is a 2-dimensional subspace. Every Jacobi field can be uniquely decomposed into the sum of a tangential Jacobi field plus a normal Jacobi field, just by decomposing its initial value and initial derivative.

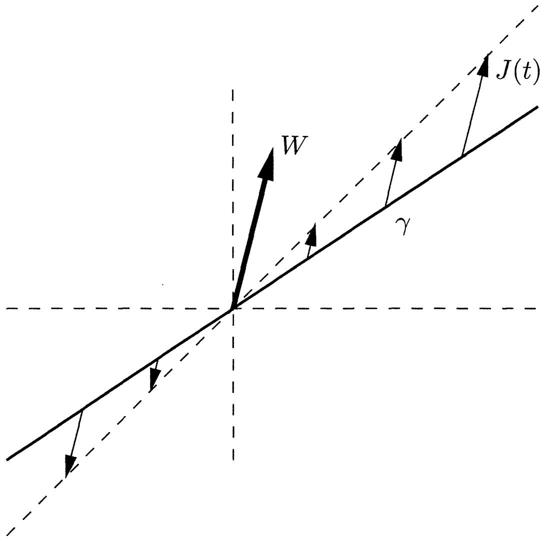


FIGURE 10.4. A Jacobi field in normal coordinates.

Computations of Jacobi Fields

In Riemannian normal coordinates, half of the Jacobi fields are easy to write down explicitly.

Lemma 10.7. *Let $p \in M$, let (x^i) be normal coordinates on a neighborhood $\mathcal{U}$ of p, and let γ be a radial geodesic starting at p. For any $W = W^i \partial_i \in T_pM$, the Jacobi field J along γ such that $J(0) = 0$ and $D_tJ(0) = W$ (see Figure 10.4) is given in normal coordinates by the formula*

$$J(t) = tW^i \partial_i. \tag{10.4}$$

Proof. An easy computation using formula (4.10) for covariant derivatives in coordinates shows that J satisfies the specified initial conditions, so it suffices to show that J is a Jacobi field. If we set $V = \dot\gamma(0) \in T_pM$, then we know from Lemma 5.11 that γ is given in coordinates by the formula $\gamma(t) = (tV^1, \ldots, tV^n)$. Now consider the variation Γ given in coordinates by

$$\Gamma(s,t) = (t(V^1 + sW^1), \ldots, t(V^n + sW^n)).$$

Again using Lemma 5.11, we see that Γ is a variation through geodesics. Therefore its variation field $\partial_s\Gamma(0,t)$ is a Jacobi field. Differentiating $\Gamma(s,t)$ with respect to s shows that its variation field is $J(t)$. □

For metrics with constant sectional curvature, we have a different kind of explicit formula for Jacobi fields—this one expresses a Jacobi field as a scalar multiple of a parallel vector field.

Lemma 10.8. *Suppose (M,g) is a Riemannian manifold with constant sectional curvature C, and γ is a unit speed geodesic in M. The normal Jacobi fields along γ vanishing at $t=0$ are precisely the vector fields*

$$J(t) = u(t)E(t), \tag{10.5}$$

where E is any parallel normal vector field along γ, and $u(t)$ is given by

$$u(t) = \begin{cases} t, & C = 0; \\ R\sin\dfrac{t}{R}, & C = \dfrac{1}{R^2} > 0; \\ R\sinh\dfrac{t}{R}, & C = -\dfrac{1}{R^2} < 0. \end{cases} \tag{10.6}$$

Proof. Since g has constant curvature, its curvature endomorphism is given by the formula of Lemma 8.10:

$$R(X,Y)Z = C\big(\langle Y,Z\rangle X - \langle X,Z\rangle Y\big).$$

Substituting this into the Jacobi equation, we find that a normal Jacobi field J satisfies

$$\begin{aligned} 0 &= D_t^2 J + C\big(\langle\dot\gamma,\dot\gamma\rangle J - \langle J,\dot\gamma\rangle\dot\gamma\big) \\ &= D_t^2 J + CJ, \end{aligned} \tag{10.7}$$

where we have used the facts that $|\dot\gamma|^2 = 1$ and $\langle J,\dot\gamma\rangle = 0$.

Since (10.7) says that the second covariant derivative of J is a multiple of J itself, it is reasonable to try to construct a solution by choosing a parallel normal vector field E along γ and setting $J(t) = u(t)E(t)$ for some function u to be determined. Plugging this into (10.7), we find that J is a Jacobi field provided u is a solution to the differential equation

$$\ddot u(t) + Cu(t) = 0.$$

It is an easy matter to solve this ODE explicitly. In particular, the solutions satisfying $u(0) = 0$ are constant multiples of the functions given in (10.6). This construction yields all the normal Jacobi fields vanishing at 0, since there is an $(n-1)$-dimensional space of them, and the space of parallel normal vector fields has the same dimension. □

Combining the formulas in the last two lemmas, we obtain our first application of Jacobi fields: explicit expressions for constant curvature metrics in normal coordinates.

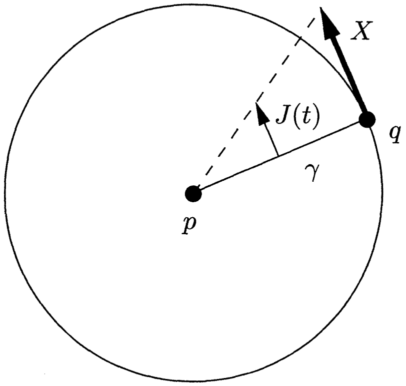


FIGURE 10.5. A vector X tangent to a geodesic sphere is the value of a normal Jacobi field.

Proposition 10.9. *Suppose (M,g) is a Riemannian manifold with constant sectional curvature C. Let (x^i) be Riemannian normal coordinates on a normal neighborhood $\mathfrak{U}$ of $p \in M$, let $|\cdot|_{\bar g}$ be the Euclidean norm in these coordinates, and let r be the radial distance function. For any $q \in \mathfrak{U} - \{p\}$ and $V \in T_qM$, write $V = V^\top + V^\perp$, where $V^\top$ is tangent to the sphere $\{r = \text{constant}\}$ through q and $V^\perp$ is a multiple of $\partial/\partial r$. The metric g can be written*

$$g(V,V) = \begin{cases} |V^\perp|^2_{\bar g} + |V^\top|^2_{\bar g}, & K = 0; \\ |V^\perp|^2_{\bar g} + \dfrac{R^2}{r^2}\left(\sin^2 \dfrac{r}{R}\right)|V^\top|^2_{\bar g}, & C = \dfrac{1}{R^2} > 0; \\ |V^\perp|^2_{\bar g} + \dfrac{R^2}{r^2}\left(\sinh^2 \dfrac{r}{R}\right)|V^\top|^2_{\bar g}, & C = -\dfrac{1}{R^2} < 0. \end{cases} \tag{10.8}$$

Proof. By the Gauss lemma, the decomposition $V = V^\top + V^\perp$ is orthogonal, so $|V|^2_g = |V^\perp|^2_g + |V^\top|^2_g$. Since $\partial/\partial r$ is a unit vector in both the g and $\bar g$ norms, it is immediate that $|V^\perp|_g = |V^\perp|_{\bar g}$. Thus we need only compute $|V^\top|_g$.

Set $X = V^\top$, and let γ denote the unit speed radial geodesic from p to q. By Lemma 10.7, X is the value of a Jacobi field J along γ that vanishes at p (Figure 10.5), namely $X = J(r)$, where $r = d(p,q)$ and

$$J(t) = \frac{t}{r} X^i \partial_i. \tag{10.9}$$

Because J is orthogonal to $\dot\gamma$ at p and q, it is normal by Lemma 10.6.

Now J can also be written in the form $J(t) = u(t)E(t)$ as in Lemma 10.8. In this representation,

$$D_t J(0) = \dot u(0)E(0) = E(0),$$

since $\dot{u}(0) = 1$ in each of the cases of (10.6). Therefore, since E is parallel and thus of constant length,

$$|X|^2 = |J(r)|^2 = |u(r)|^2|E(r)|^2 = |u(r)|^2|E(0)|^2 = |u(r)|^2|D_tJ(0)|^2. \tag{10.10}$$

Observe that $D_tJ(0) = (1/r)X^i\partial_i|_p$ by (10.9). Since g agrees with $\bar{g}$ at p, we have

$$|D_tJ(0)| = \frac{1}{r}\left|X^i\partial_i\big|_p\right|_g = \frac{1}{r}|X|_{\bar{g}}.$$

Inserting this into (10.10) and using formula (10.6) for $u(r)$ completes the proof. □

Proposition 10.10. (Local Uniqueness of Constant Curvature Metrics) *Let (M,g) and $(\widetilde{M},\tilde{g})$ be Riemannian manifolds with constant sectional curvature C. For any points $p \in M$, $\tilde{p} \in \widetilde{M}$, there exist neighborhoods $\mathcal{U}$ of p and $\widetilde{\mathcal{U}}$ of $\tilde{p}$ and an isometry $F\colon \mathcal{U} \to \widetilde{\mathcal{U}}$.*

Proof. Choose $p \in M$ and $\tilde{p} \in \widetilde{M}$, and let $\mathcal{U}$ and $\widetilde{\mathcal{U}}$ be geodesic balls of small radius ε around p and $\tilde{p}$, respectively. Riemannian normal coordinates give maps $\varphi\colon \mathcal{U} \to B_\varepsilon(0) \subset \mathbf{R}^n$ and $\widetilde{\varphi}\colon \widetilde{\mathcal{U}} \to B_\varepsilon(0) \subset \mathbf{R}^n$, under which both metrics are given by (10.8) (Figure 10.6). Therefore $\widetilde{\varphi}^{-1} \circ \varphi$ is the required local isometry. □

Conjugate Points

Our next application of Jacobi fields is to study the question of when the exponential map is a local diffeomorphism. If (M,g) is complete, we know that $\exp_p$ is defined on all of T_pM, and is a local diffeomorphism near 0. However, it may well happen that it ceases to be even a local diffeomorphism at points far away.

An enlightening example is provided by the sphere $\mathbf{S}^n_R$. All geodesics starting at a given point p meet at the antipodal point, which is at a distance of πR along each geodesic. The exponential map is a diffeomorphism on the ball $B_{\pi R}(0)$, but it fails to be a local diffeomorphism at all points on the sphere of radius πR in $T_p\mathbf{S}^n_R$ (Figure 10.7). Moreover, Lemma 10.8 shows that each Jacobi field on $\mathbf{S}^n_R$ vanishing at p has its first zero precisely at distance πR.

On the other hand, formula (10.4) shows that if $\mathcal{U}$ is a normal neighborhood of p (the image of a set on which $\exp_p$ is a diffeomorphism), no Jacobi field that vanishes at p can vanish at any other point in $\mathcal{U}$. We might thus be led to expect a relationship between zeros of Jacobi fields

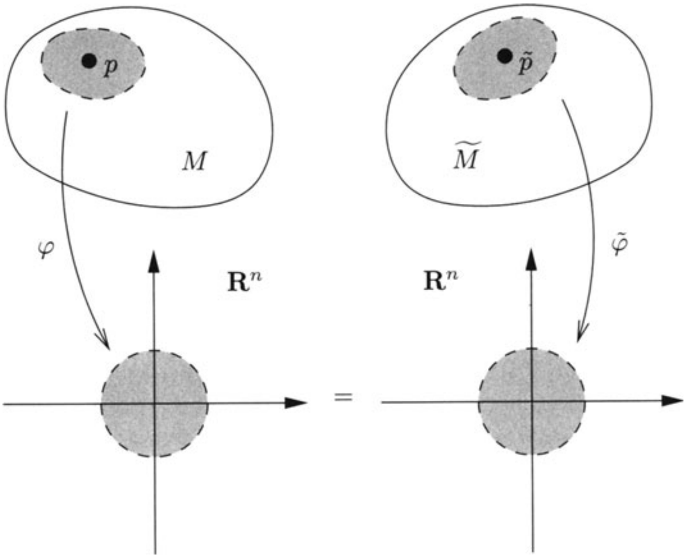


FIGURE 10.6. Local isometry constructed from normal coordinate charts.

and singularities of the exponential map (i.e., points where it fails to be a local diffeomorphism).

If γ is a geodesic segment joining $p, q \in M$, q is said to be *conjugate to p along γ* if there is a Jacobi field along γ vanishing at p and q but not identically zero (Figure 10.8). The *order* or *multiplicity* of conjugacy is the dimension of the space of Jacobi fields vanishing at p and q. From the existence and uniqueness theorem for Jacobi fields, there is an n-dimensional space of Jacobi fields that vanish at p; since tangential Jacobi fields vanish at most at one point, the order of conjugacy of two points p and q can be at most $n-1$. This bound is sharp: Lemma 10.8 shows that if p and q are antipodal points on $\mathbf{S}^n_R$, there is a Jacobi field vanishing at p and q for each parallel normal vector field along γ; thus in that case p and q are conjugate to order exactly $n-1$.

The most important fact about conjugate points is that they are precisely the images of singularities of the exponential map, as the following proposition shows.

Proposition 10.11. *Suppose $p \in M$, $V \in T_pM$, and $q = \exp_p V$. Then $\exp_p$ is a local diffeomorphism in a neighborhood of V if and only if q is not conjugate to p along the geodesic $\gamma(t) = \exp_p tV$, $t \in [0,1]$.*

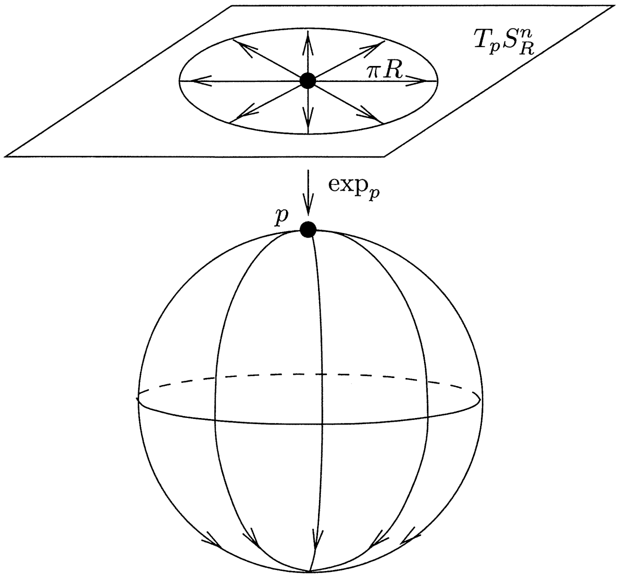


FIGURE 10.7. The exponential map of the sphere.

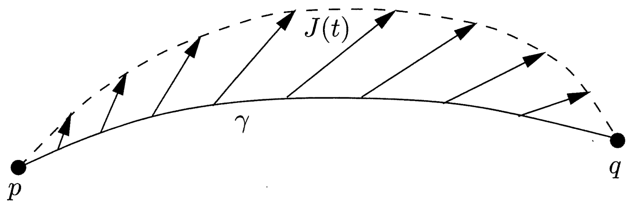


FIGURE 10.8. Conjugate points.

Proof. By the inverse function theorem, $\exp_p$ is a local diffeomorphism near V if and only if $(\exp_p)_*$ is an isomorphism at V, and by dimensional considerations, this occurs if and only if $(\exp_p)_*$ is injective at V.

Identifying $T_V(T_pM)$ with T_pM as usual, we can compute the push-forward $(\exp_p)_*$ at V as follows:

$$(\exp_p)_*W = \left.\frac{d}{ds}\right|_{s=0} \exp_p(V + sW).$$

To compute this, we define a variation of γ through geodesics (Figure 10.9) by

$$\Gamma_W(s,t) = \exp_p t(V + sW).$$

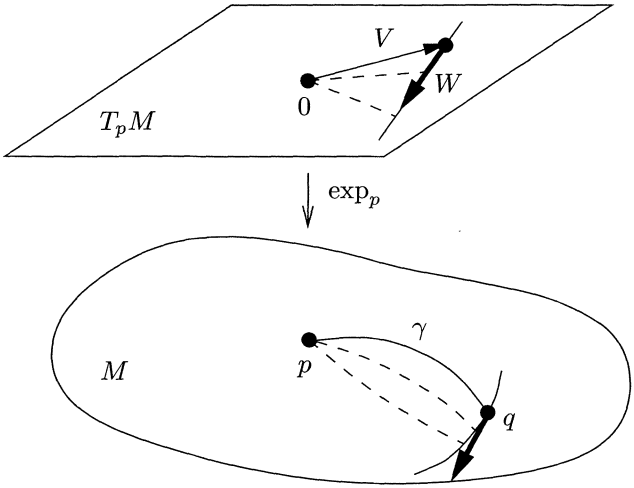


FIGURE 10.9. Computing $(\exp_p)_*W$.

Then the variation field $J_W(t) = \partial_s\Gamma_W(0,t)$ is a Jacobi field along γ, and

$$J_W(1) = (\exp_p)_*W.$$

Since $W \in T_pM$ is arbitrary, there is an n-dimensional space of such Jacobi fields, and so these are all the Jacobi fields along γ that vanish at p. (If γ is contained in a normal neighborhood, these are just the Jacobi fields of the form (10.4) in normal coordinates.)

Therefore, $(\exp_p)_*$ fails to be an isomorphism at V when there is a vector W such that $(exp_p)_*W = 0$, which occurs precisely when there is a Jacobi field J_W along γ with $J_W(0) = J_W(q) = 0$. □

As Proposition 10.4 shows, the "natural" way to specify a unique Jacobi field is by giving its initial value and initial derivative. However, in a number of the arguments above, we have had to construct Jacobi fields along a geodesic γ satisfying $J(0) = 0$ and $J(b) = W$ for some specific vector W. More generally, one can pose the *two-point boundary problem* for Jacobi fields: Given $V \in T_{\gamma(a)}M$ and $W \in T_{\gamma(b)}M$, find a Jacobi field J along γ such that $J(a) = V$ and $J(b) = W$. Another interesting property of conjugate points is that they are the obstruction to solving the two-point boundary problem, as the next exercise shows.

Exercise 10.2. Suppose $\gamma\colon [a,b] \to M$ is a geodesic. Show that the two-point boundary problem for Jacobi fields is uniquely solvable for every pair of vectors $V \in T_{\gamma(a)}M$ and $W \in T_{\gamma(b)}M$ if and only if $\gamma(a)$ and $\gamma(b)$ are not conjugate along γ.

The Second Variation Formula

Our last task in this chapter is to study the question of which geodesics are minimizing. In our proof that any minimizing curve is a geodesic, we imitated the first-derivative test of elementary calculus: If a geodesic γ is minimizing, then the first derivative of the length functional must vanish for any proper variation of γ. Now we imitate the second-derivative test: If γ is minimizing, the second derivative must be nonnegative. First, we must compute this second derivative. In keeping with classical terminology, we call it the *second variation* of the length functional.

Theorem 10.12. (The Second Variation Formula) *Let* $\gamma\colon [a,b] \to M$ *be a unit speed geodesic,* Γ *a proper variation of* γ*, and* V *its variation field. The second variation of* $L(\Gamma_s)$ *is given by the following formula:*

$$\left.\frac{d^2}{ds^2}\right|_{s=0} L(\Gamma_s) = \int_a^b \left(\left|D_t V^\perp\right|^2 - Rm(V^\perp, \dot\gamma, \dot\gamma, V^\perp) \right) dt, \qquad (10.11)$$

where $V^\perp$ *is the normal component of* V.

Proof. As usual, write $T = \partial_t \Gamma$ and $S = \partial_s \Gamma$. We begin, as we did when computing the first variation formula, by restricting to a rectangle $(-\varepsilon, \varepsilon) \times [a_{i-1}, a_i]$ where Γ is smooth. From (6.3) we have, for any s,

$$\frac{d}{ds} L(\Gamma_s|_{[a_{i-1},a_i]}) = \int_{a_{i-1}}^{a_i} \frac{\langle D_t S, T\rangle}{\langle T, T\rangle^{1/2}}\, dt.$$

Differentiating again with respect to s, and using the symmetry lemma and Lemma 10.1,

$$\begin{aligned}
&\frac{d^2}{ds^2} L(\Gamma_s|_{[a_{i-1},a_i]}) \\
&= \int_{a_{i-1}}^{a_i} \left(\frac{\langle D_s D_t S, T\rangle}{\langle T,T\rangle^{1/2}} + \frac{\langle D_t S, D_s T\rangle}{\langle T,T\rangle^{1/2}} - \frac{1}{2} \frac{\langle D_t S, T\rangle 2 \langle D_s T, T\rangle}{\langle T, T\rangle^{3/2}} \right) dt \\
&= \int_{a_{i-1}}^{a_i} \left(\frac{\langle D_t D_s S + R(S,T)S, T\rangle}{|T|} + \frac{\langle D_t S, D_t S\rangle}{|T|} - \frac{\langle D_t S, T\rangle^2}{|T|^3} \right) dt.
\end{aligned}$$

Now restrict to $s = 0$, where $|T| = 1$:

$$\begin{aligned}
\left.\frac{d^2}{ds^2}\right|_{s=0} L(\Gamma_s|_{[a_{i-1},a_i]}) = \int_{a_{i-1}}^{a_i} \big(\langle D_t D_s S, T\rangle - Rm(S,T,T,S) \\
+ |D_t S|^2 - \langle D_t S, T\rangle^2 \big)\, dt \Big|_{s=0}.
\end{aligned} \qquad (10.12)$$

Because $D_tT = D_t\dot\gamma = 0$ when $s = 0$, the first term in (10.12) can be integrated as follows:

$$\begin{aligned} \int_{a_{i-1}}^{a_i} \langle D_tD_sS, T\rangle dt &= \int_{a_{i-1}}^{a_i} \frac{\partial}{\partial t}\langle D_sS, T\rangle dt \\ &= \langle D_sS, T\rangle \Big|_{t=a_{i-1}}^{t=a_i}. \end{aligned} \tag{10.13}$$

Notice that $S(s,t) = 0$ for all s at the endpoints $t = a_0 = a$ and $t = a_k = b$ because Γ is a proper variation, so $D_sS = 0$ there. Moreover, along the boundaries $\{t = a_i\}$ of the smooth regions, $D_sS = D_s(\partial_s\Gamma)$ depends only on the values of Γ when $t = a_i$, and it is smooth up to the line $\{t = a_i\}$ from both sides; therefore D_sS is continuous for all (s,t). Thus when we insert (10.13) into (10.12) and sum over i, the boundary contributions from the first term all cancel, and we get

$$\begin{aligned} \frac{d^2}{ds^2}\Big|_{s=0} L(\Gamma_s) &= \int_a^b \left(|D_tS|^2 - \langle D_tS, T\rangle^2 - Rm(S,T,T,S)\right) dt\Big|_{s=0} \\ &= \int_a^b \left(|D_tV|^2 - \langle D_tV, \dot\gamma\rangle^2 - Rm(V,\dot\gamma,\dot\gamma,V)\right) dt. \end{aligned} \tag{10.14}$$

Any vector field V along γ can be written uniquely as $V = V^\top + V^\perp$, where $V^\top$ is tangential and $V^\perp$ is normal. Explicitly,

$$V^\top = \langle V, \dot\gamma\rangle\dot\gamma; \qquad V^\perp = V - V^\top.$$

Because $D_t\dot\gamma = 0$, it follows that

$$D_tV^\top = \langle D_tV, \dot\gamma\rangle\dot\gamma = (D_tV)^\top; \qquad D_tV^\perp = (D_tV)^\perp.$$

Therefore,

$$|D_tV|^2 = |(D_tV)^\top|^2 + |(D_tV)^\perp|^2 = \langle D_tV, \dot\gamma\rangle^2 + |D_tV^\perp|^2.$$

Also,

$$Rm(V,\dot\gamma,\dot\gamma,V) = Rm(V^\perp,\dot\gamma,\dot\gamma,V^\perp)$$

because $Rm(\dot\gamma,\dot\gamma,\cdot,\cdot) = Rm(\cdot,\cdot,\dot\gamma,\dot\gamma) = 0$. Substituting these relations into (10.14) gives (10.11). □

It should come as no surprise that the second variation depends only on the normal component of V; intuitively, the tangential component of V contributes only to a reparametrization of γ, and length is independent of

parametrization. For this reason, we generally apply the second variation formula only to variations whose variation fields are proper and normal.

We define a symmetric bilinear form I, called the *index form,* on the space of proper normal vector fields along γ by

$$I(V,W) = \int_a^b (\langle D_tV, D_tW\rangle - Rm(V,\dot\gamma,\dot\gamma,W))\,dt. \tag{10.15}$$

You should think of $I(V,W)$ as a sort of "Hessian" or second derivative of the length functional. Because every proper normal vector field along γ is the variation field of some proper variation, the preceding theorem can be rephrased in terms of the index form in the following way.

Corollary 10.13. *If Γ is a proper variation of a unit speed geodesic γ whose variation field is a proper normal vector field V, the second variation of $L(\Gamma_s)$ is $I(V,V)$. In particular, if γ is minimizing, then $I(V,V)\geq 0$ for any proper normal vector field along γ.*

The next proposition gives another expression for I, which makes the role of the Jacobi equation more evident.

Proposition 10.14. *For any pair of proper normal vector fields V,W along a geodesic segment γ,*

$$I(V,W) = -\int_a^b \left\langle D_t^2V + R(V,\dot\gamma)\dot\gamma, W\right\rangle dt - \sum_{i=1}^k \langle \Delta_i D_tV, W(a_i)\rangle\,, \tag{10.16}$$

where $\{a_i\}$ are the points where V is not smooth, and $\Delta_i D_tV$ is the jump in D_tV at $t=a_i$.

Proof. On any subinterval $[a_{i-1},a_i]$ where V and W are smooth,

$$\frac{d}{dt}\langle D_tV, W\rangle = \left\langle D_t^2V, W\right\rangle + \langle D_tV, D_tW\rangle\,.$$

Thus, by the fundamental theorem of calculus,

$$\int_{a_{i-1}}^{a_i} \langle D_tV, D_tW\rangle\,dt = -\int_{a_{i-1}}^{a_i} \left\langle D_t^2V, W\right\rangle + \left.\langle D_tV, W\rangle\right|_{a_{i-1}}^{a_i}.$$

Summing over i, and noting that W is continuous at $t=a_i$ and $W(a) = W(b) = 0$, we get (10.16). □

Geodesics Do Not Minimize Past Conjugate Points

In this section, we use the second variation to prove another extremely important fact about conjugate points: No geodesic is minimizing past its

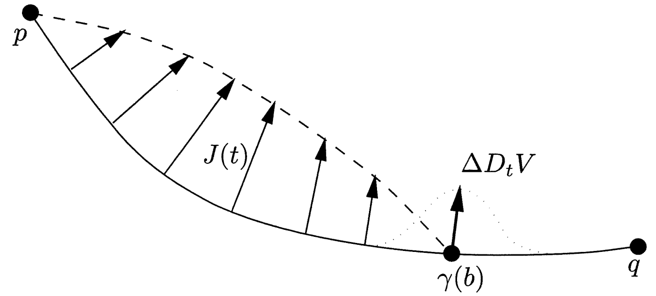


FIGURE 10.10. Constructing a vector field X with $I(X,X) < 0$.

first conjugate point. The geometric intuition is as follows: Suppose γ is minimizing. If $q = \gamma(b)$ is conjugate to $p = \gamma(a)$ along γ, and J is a Jacobi field vanishing at p and q, there is a variation of γ through geodesics, all of which start at p. Since $J(q) = 0$, we can expect them to end "almost" at q. If they really did all end at q, we could construct a broken geodesic by following some Γ_s from p to q and then following γ from q to $\gamma(b+\varepsilon)$, which would have the same length and thus would also be a minimizing curve. But this is impossible: as the proof of Theorem 6.6 shows, a broken geodesic can always be shortened by rounding the corner.

The problem with this heuristic argument is that there is no guarantee that we can construct a variation through geodesics that actually end at q. The proof of the following theorem is based on an "infinitesimal" version of rounding the corner to obtain a shorter curve.

Theorem 10.15. *If γ is a geodesic segment from p to q that has an interior conjugate point to p, then there exists a proper normal vector field X along γ such that $I(X,X) < 0$. In particular, γ is not minimizing.*

Proof. Suppose $\gamma\colon [0,b] \to M$ is a unit speed parametrization of γ, and $\gamma(a)$ is conjugate to $\gamma(0)$ for some $0 < a < b$. This means there is a nontrivial normal Jacobi field J along $\gamma|_{[0,a]}$ that vanishes at $t = 0$ and $t = a$. Define a vector field V along all of γ by

$$V(t) = \begin{cases} J(t), & t \in [0,a]; \\ 0, & t \in [a,b]. \end{cases}$$

This is a proper, normal, piecewise smooth vector field along γ.

Let W be a smooth proper normal vector field along γ such that $W(b)$ is equal to the jump $\Delta D_t V$ at $t = b$ (Figure 10.10). Such a vector field is easily constructed in local coordinates and extended to all of γ by a bump function. Note that $\Delta D_t V = -D_t J(b)$ is not zero, because otherwise J would be a Jacobi field satisfying $J(b) = D_t J(b) = 0$, and thus would be identically zero.

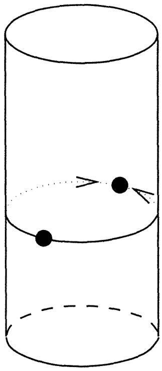


FIGURE 10.11. Geodesics on the cylinder.

For small positive ε, let $X_\varepsilon = V + \varepsilon W$. Then

$$\begin{aligned} I(X_\varepsilon, X_\varepsilon) &= I(V + \varepsilon W, V + \varepsilon W) \\ &= I(V, V) + 2\varepsilon I(V, W) + \varepsilon^2 I(W, W). \end{aligned}$$

Since V satisfies the Jacobi equation on each subinterval $[0, a]$ and $[a, b]$, and $V(a) = 0$, (10.16) gives

$$I(V, V) = -\langle \Delta D_t V, V(a) \rangle = 0.$$

Similarly,

$$I(V, W) = -\langle \Delta D_t V, W(b) \rangle = -|W(b)|^2.$$

Thus

$$I(X_\varepsilon, X_\varepsilon) = -2\varepsilon |W(b)|^2 + \varepsilon^2 I(W, W).$$

If we choose ε small enough, this is strictly negative. □

There is a far-reaching quantitative generalization of Theorem 10.15 called the Morse index theorem, which we do not treat here. The *index* of a geodesic segment is defined to be the maximum dimension of a linear space of proper normal vector fields on which I is negative definite. Roughly speaking, the index is the number of independent directions in which γ can be deformed to decrease its length. (Analogously, the index of a critical point of a function on $\mathbf{R}^n$ is defined as the number of negative eigenvalues of its Hessian.) The Morse index theorem says that the index of any geodesic segment is finite, and is equal to the number of its interior conjugate points counted with multiplicity. (Proofs can be found in [CE75], [dC92], or [Spi79, volume 4].)

It is important to note, by the way, that the converse of Theorem 10.15 is not true: a geodesic without conjugate points need not be minimizing. For example, on the cylinder $\mathbf{S}^1 \times \mathbf{R}$, there are no conjugate points along any geodesic; but no geodesic that wraps more than halfway around the cylinder is minimizing (Figure 10.11). Therefore it is useful to make the following definitions. Suppose γ is a geodesic starting at p. Let $B = \sup\{b > 0 : \gamma|_{[0,b]} \text{ is minimizing}\}$. If $B < \infty$, we call $q = \gamma(B)$ the *cut point* of p along γ. The *cut locus* of p is the set of all points $q \in M$ such that q is the cut point of p along some geodesic. (Analogously, the *conjugate locus* of p is the set of points q such that q is the first conjugate point to p along some geodesic.) The preceding theorem can be interpreted as saying that the cut point (if it exists) occurs at or before the first conjugate point along any geodesic.

Problems

10-1. Extend the result of Lemma 10.8 by finding a formula for *all* normal Jacobi fields in the constant curvature case, not just the ones that vanish at 0.

10-2. Suppose that all sectional curvatures of M are nonpositive. Use the results of this chapter to show that the conjugate locus of any point is empty. [We will give a more geometric proof in the next chapter.]

10-3. Suppose (M, g) is a Riemannian manifold and $p \in M$. Show that the second-order Taylor series of g in normal coordinates centered at p is

$$g_{ij}(x) = \delta_{ij} - \frac{1}{3}\sum_{kl} R_{iklj}x^k x^l + O(|x|^3).$$

[Hint: Let $\gamma(t) = (tV^1, \ldots, tV^n)$ be a radial geodesic and $J(t) = tW^i\partial_i$ a Jacobi field along γ, and compute the first four t-derivatives of $|J(t)|^2$ at $t = 0$ in two ways.]

11
Curvature and Topology

In this final chapter, we bring together most of the tools we have developed so far to prove some significant local-global theorems relating curvature and topology. Before treating the topological theorems themselves, we prove some comparison theorems for manifolds whose curvature is bounded above. These comparisons are based on a simple ODE comparison theorem due to Sturm, and show that if the curvature is bounded above by a constant, then the metric in normal coordinates is bounded below by the corresponding constant curvature metric.

We then state and prove several of the most important local-global theorems of Riemannian geometry. The first one, the Cartan–Hadamard theorem, topologically characterizes complete, simply-connected manifolds with nonpositive sectional curvature: they are all diffeomorphic to $\mathbf{R}^n$. The second, Bonnet's theorem, says that a complete manifold with sectional curvatures bounded below by a positive constant must be compact and have a finite fundamental group; a generalization called Myers's theorem allows positive sectional curvature to be replaced by positive Ricci curvature. The last theorem in this chapter says that complete manifolds with constant sectional curvature are all quotients of the model spaces by discrete subgroups of their isometry groups.

Some Comparison Theorems

We begin this chapter by proving that an upper bound on sectional curvature produces a lower bound on Jacobi fields, on the distance to conjugate points, and on the metric in normal coordinates. Our starting point is the following very classical comparison theorem for ordinary differential equations. This result can be found in various guises in the literature (cf., for example, [Spi79, volume 4] or [BR78, Theorem II.6]), but all are essentially equivalent to the one presented here.

Theorem 11.1. (Sturm Comparison Theorem) *Suppose u and v are differentiable real-valued functions on $[0,T]$, twice differentiable on $(0,T)$, and $u > 0$ on $(0,T)$. Suppose further that u and v satisfy*

$$\begin{gathered} \ddot{u}(t) + a(t)u(t) = 0 \\ \ddot{v}(t) + a(t)v(t) \geq 0 \\ u(0) = v(0) = 0, \qquad \dot{u}(0) = \dot{v}(0) > 0 \end{gathered}$$

for some function $a\colon [0,T] \to \mathbf{R}$. Then $v(t) \geq u(t)$ on $[0,T]$.

Proof. Consider the function $f(t) = v(t)/u(t)$ defined on $(0,T)$. It follows from l'Hôpital's rule that $\lim_{t\to 0} f(t) = \dot{v}(0)/\dot{u}(0) = 1$. Since f is differentiable on $(0,T)$, if we could show that $\dot{f} \geq 0$ there it would follow from elementary calculus that $f \geq 1$ and therefore $v \geq u$ on $(0,T)$, and by continuity also on $[0,T]$. Differentiating,

$$\frac{d}{dt}\left(\frac{v}{u}\right) = \frac{\dot{v}u - v\dot{u}}{u^2}.$$

Thus to show $\dot{f} \geq 0$ it would suffice to show $\dot{v}u - v\dot{u} \geq 0$. Since $\dot{v}(0)u(0) - v(0)\dot{u}(0) = 0$, we need only show this expression has nonnegative derivative. Differentiating again and substituting the ODE for u,

$$\frac{d}{dt}(\dot{v}u - v\dot{u}) = \ddot{v}u + \dot{v}\dot{u} - \dot{v}\dot{u} - v\ddot{u} = \ddot{v}u + avu \geq 0.$$

This proves the theorem. □

Theorem 11.2. (Jacobi Field Comparison Theorem) *Suppose (M,g) is a Riemannian manifold with all sectional curvatures bounded above by a constant C. If γ is a unit speed geodesic in M, and J is any normal Jacobi field along γ such that $J(0) = 0$, then*

$$|J(t)| \geq \begin{cases} t\,|D_tJ(0)| & \textit{for } 0 \leq t, & \textit{if } C = 0; \\ R\sin\dfrac{t}{R}\,|D_tJ(0)| & \textit{for } 0 \leq t \leq \pi R, & \textit{if } C = \dfrac{1}{R^2} > 0; \\ R\sinh\dfrac{t}{R}\,|D_tJ(0)| & \textit{for } 0 \leq t, & \textit{if } C = -\dfrac{1}{R^2} < 0. \end{cases}$$

Proof. The function $|J(t)|$ is smooth wherever $J(t) \neq 0$. Using the Jacobi equation, we compute

$$\begin{aligned}\frac{d^2}{dt^2}|J| &= \frac{d}{dt}\frac{\langle D_tJ, J\rangle}{\langle J, J\rangle^{1/2}} \\ &= \frac{\langle D_t^2J, J\rangle}{\langle J, J\rangle^{1/2}} + \frac{\langle D_tJ, D_tJ\rangle}{\langle J, J\rangle^{1/2}} - \frac{\langle D_tJ, J\rangle^2}{\langle J, J\rangle^{3/2}} \\ &= -\frac{\langle R(J,\dot\gamma)\dot\gamma, J\rangle}{|J|} + \frac{|D_tJ|^2}{|J|} - \frac{\langle D_tJ, J\rangle^2}{|J|^3}.\end{aligned}$$

By the Schwartz inequality, $\langle D_tJ, J\rangle^2 \leq |D_tJ|^2|J|^2$, so the sum of the last two terms above is nonnegative. Thus

$$\frac{d^2}{dt^2}|J| \geq -\frac{Rm(J,\dot\gamma,\dot\gamma,J)}{|J|}.$$

Since $\langle J, \dot\gamma\rangle = 0$ and $|\dot\gamma| = 1$, $Rm(J,\dot\gamma,\dot\gamma,J)/|J|^2$ is the sectional curvature of the plane spanned by J and $\dot\gamma$. Therefore our assumption on the sectional curvatures of M guarantees that $Rm(J,\dot\gamma,\dot\gamma,J)/|J|^2 \leq C$, so $|J|$ satisfies the differential inequality

$$\frac{d^2}{dt^2}|J| \geq -C|J|$$

wherever $|J| > 0$.

We wish to use the Sturm comparison theorem to compare $|J|$ with the solution u to $\ddot u + Cu = 0$ given by (10.6). To do so, we need to arrange that $d|J|/dt = 1$ at $t = 0$, because $\dot u(0) = 1$. Multiplying J by a positive constant, we may assume without loss of generality that $|D_tJ(0)| = 1$. From Lemma 10.7, J can be written near $t = 0$ as $J(t) = tW(t)$, where W is a smooth vector field. (It is the one given in normal coordinates by $W(t) = W^i\partial_i$ for some constants $W^1, \dots, W^n$, but that is irrelevant here.) Therefore,

$$\begin{aligned}\left.\frac{d}{dt}\right|_{t=0}|J(t)| &= \lim_{t\to0}\frac{|J(t)| - |J(0)|}{t} \\ &= \lim_{t\to0}\frac{t|W(t)|}{t} = |W(0)| = |D_tJ(0)| = 1.\end{aligned}$$

Now the Sturm comparison theorem applies to show that $|J| \geq u$, provided $|J|$ is nonzero (to ensure that it is smooth). The fact that $d|J|/dt = 1$ at $t = 0$ means $|J| > 0$ on some interval $(0, \varepsilon)$, and $|J|$ cannot attain its first zero before u does without contradicting the estimate $|J| \geq u$. Thus $|J| \geq u$ as long as $u \geq 0$, which proves the theorem. □

Corollary 11.3. (Conjugate Point Comparison Theorem) *Suppose all sectional curvatures of (M, g) are bounded above by a constant C. If*

$C \leq 0$, then no point of M has conjugate points along any geodesic. If $C = 1/R^2 > 0$, then the first conjugate point along any geodesic occurs at a distance of at least πR.

Proof. If $C \leq 0$, the Jacobi field comparison theorem implies that any nontrivial normal Jacobi field vanishing at $t = 0$ satisfies $|J(t)| > 0$ for all $t > 0$. Similarly, if $C > 0$, then $|J(t)| \geq (\text{constant})\sin(t/R) > 0$ for $0 < t < \pi R$. □

Corollary 11.4. (Metric Comparison Theorem) *Suppose all sectional curvatures of (M, g) are bounded above by a constant C. In any normal coordinate chart, $g(V, V) \geq g_C(V, V)$, where g_C is the constant curvature metric given by formula* (10.8).

Proof. Decomposing a vector V into components $V^\top$ tangent to the geodesic sphere and $V^\perp$ tangent to the radial geodesics as in the proof of Proposition 10.9 gives

$$g(V, V) = g(V^\perp, V^\perp) + g(V^\top, V^\top).$$

Just as in that proof, $g(V^\perp, V^\perp) = \bar{g}(V^\perp, V^\perp) = g_C(V^\perp, V^\perp)$. Also, $V^\top$ is the value of some normal Jacobi field vanishing at $t = 0$, so the Jacobi field comparison theorem gives $g(V^\top, V^\top) \geq g_C(V^\top, V^\top)$. □

The general information provided by these results is that a nonpositive upper bound on curvature forces geodesics to "spread out," while a positive upper bound prevents them from converging too fast.

Manifolds of Negative Curvature

Our first major local-global theorem in arbitrary dimensions is the following characterization of simply-connected manifolds of nonpositive sectional curvature.

Theorem 11.5. (The Cartan–Hadamard Theorem) *If M is a complete, connected manifold all of whose sectional curvatures are nonpositive, then for any point $p \in M$, $\exp_p \colon T_pM \to M$ is a covering map. In particular, the universal covering space of M is diffeomorphic to $\mathbf{R}^n$. If M is simply connected, then M itself is diffeomorphic to $\mathbf{R}^n$.*

Proof. The assumption of nonpositive curvature guarantees that p has no conjugate points along any geodesic, which can be shown by using either the conjugate point comparison theorem above or Problem 10-2. Therefore, by Proposition 10.11, $\exp_p$ is a local diffeomorphism on all of T_pM.

Let $\tilde{g}$ be the (variable-coefficient) 2-tensor field $\exp_p^* g$ defined on T_pM. Because $\exp_p^*$ is everywhere nonsingular, $\tilde{g}$ is a Riemannian metric, and

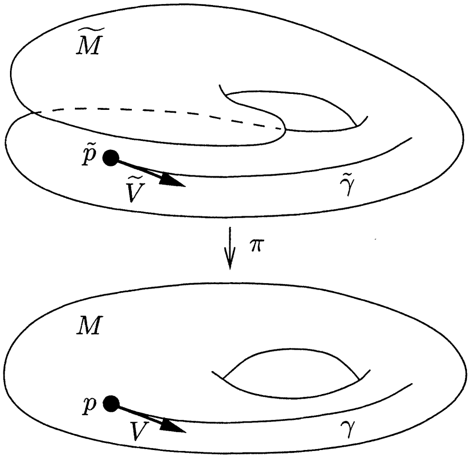


FIGURE 11.1. Lifting geodesics.

$\exp_p\colon (T_pM, \tilde{g}) \to (M, g)$ is a local isometry. It then follows from Lemma 11.6 below that $\exp_p$ is a covering map. The remaining statements of the theorem follow immediately from uniqueness of the universal covering space. □

Lemma 11.6. *Suppose $\widetilde{M}$ and M are connected Riemannian manifolds, with $\widetilde{M}$ complete, and $\pi\colon \widetilde{M} \to M$ is a local isometry. Then M is complete and π is a covering map.*

Proof. A fundamental property of covering maps is the *path-lifting property:* any continuous path γ in M lifts to a path $\tilde{\gamma}$ in $\widetilde{M}$ such that $\pi \circ \tilde{\gamma} = \gamma$. We begin by proving that π possesses the path-lifting property for geodesics: If $p \in M$, $\tilde{p} \in \pi^{-1}(p)$, and $\gamma\colon I \to M$ is a geodesic starting at p, then γ has a unique lift starting at $\tilde{p}$ (Figure 11.1). The lifted curve is necessarily also a geodesic because π is a local isometry.

To prove the path-lifting property for geodesics, let $V = \dot{\gamma}(0)$ and $\widetilde{V} = \pi_*^{-1}\dot{\gamma}(0) \in T_{\tilde{p}}\widetilde{M}$ (which is well defined because π_* is an isomorphism at each point), and let $\tilde{\gamma}$ be the geodesic in $\widetilde{M}$ with initial point $\tilde{p}$ and initial velocity $\widetilde{V}$. Because $\widetilde{M}$ is complete, $\tilde{\gamma}$ is defined for all time. Since π is a local isometry, it takes geodesics to geodesics; and since by construction $\pi(\tilde{\gamma}(0)) = \gamma(0)$ and $\pi_*\dot{\tilde{\gamma}}(0) = \dot{\gamma}(0)$, we must have $\pi \circ \tilde{\gamma} = \gamma$ on I. In particular, $\pi \circ \tilde{\gamma}$ is a geodesic defined for all t that coincides with γ on I, so γ extends to all of $\mathbf{R}$ and thus M is complete.

Next we show that π is surjective. Choose some point $\tilde{p} \in \widetilde{M}$, write $p = \pi(\tilde{p})$, and let $q \in M$ be arbitrary. Because M is connected and complete, there is a minimizing geodesic segment γ from p to q. Letting $\tilde{\gamma}$ be the lift

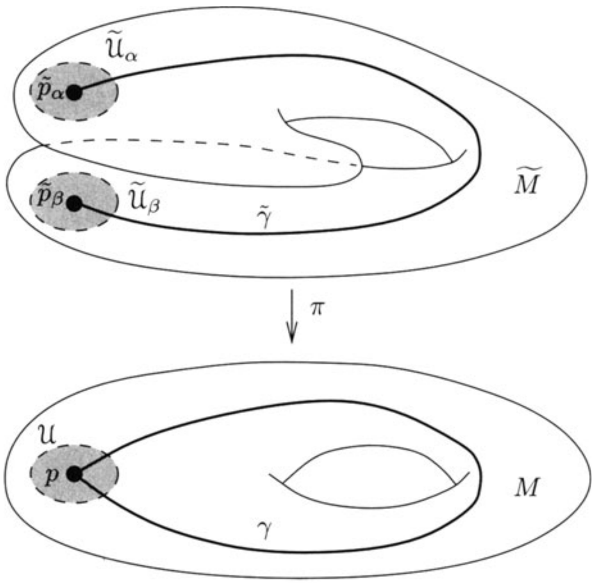


FIGURE 11.2. Proof that $\widetilde{\mathcal{U}}_\alpha$ and $\widetilde{\mathcal{U}}_\beta$ are disjoint.

of γ starting at $\tilde p$ and $r = d(p,q)$, we have $\pi(\tilde\gamma(r)) = \gamma(r) = q$, so q is in the image of π.

To show that π is a covering map, we need to show that every point $p \in M$ has a neighborhood $\mathcal{U}$ that is evenly covered, which means that $\pi^{-1}(\mathcal{U})$ is the disjoint union of open sets $\widetilde{\mathcal{U}}_\alpha$ such that $\pi\colon \widetilde{\mathcal{U}}_\alpha \to \mathcal{U}$ is a diffeomorphism. We will show, in fact, that any geodesic ball $\mathcal{U} = B_\varepsilon(p)$ is evenly covered.

Let $\pi^{-1}(p) = \{\tilde p_\alpha\}$, and for each α let $\widetilde{\mathcal{U}}_\alpha$ denote the *metric* ball of radius ε around $\tilde p_\alpha$ (we are not claiming that $\widetilde{\mathcal{U}}_\alpha$ is a geodesic ball). The first step is to show that the various sets $\widetilde{\mathcal{U}}_\alpha$ are disjoint. For any $\alpha \neq \beta$, there is a minimizing geodesic $\tilde\gamma$ from $\tilde p_\alpha$ to $\tilde p_\beta$ because $\widetilde M$ is complete. The projected curve $\gamma := \pi \circ \tilde\gamma$ is a geodesic from p to p (Figure 11.2). Such a geodesic must leave $\mathcal{U}$ and re-enter it (since all geodesics passing through p and lying in $\mathcal{U}$ are radial line segments), and thus must have length at least 2ε. This means $d(\tilde p_\alpha, \tilde p_\beta) > 2\varepsilon$, and thus by the triangle inequality $\widetilde{\mathcal{U}}_\alpha \cap \widetilde{\mathcal{U}}_\beta = \emptyset$.

The next step is to show that $\pi^{-1}(\mathcal{U}) = \bigcup_\alpha \widetilde{\mathcal{U}}_\alpha$. Since π is an isometry, it clearly maps $\widetilde{\mathcal{U}}_\alpha$ into $\mathcal{U}$. Thus we need only show $\pi^{-1}(\mathcal{U}) \subset \bigcup_\alpha \widetilde{\mathcal{U}}_\alpha$. Let $\tilde q \in \pi^{-1}(\mathcal{U})$. This means that $q := \pi(\tilde q) \in \mathcal{U}$, so there is a minimizing geodesic γ in $\mathcal{U}$ from q to p, and $r = d(q,p) < \varepsilon$. Letting $\tilde\gamma$ be the lift of γ starting at $\tilde q$, it follows that $\pi(\tilde\gamma(r)) = \gamma(r) = p$ (Figure 11.3). Therefore $\tilde\gamma(r) = \tilde p_\alpha$ for some α, and $d(\tilde q, \tilde p_\alpha) \le L(\tilde\gamma) = r < \varepsilon$, so $\tilde q \in \widetilde{\mathcal{U}}_\alpha$.

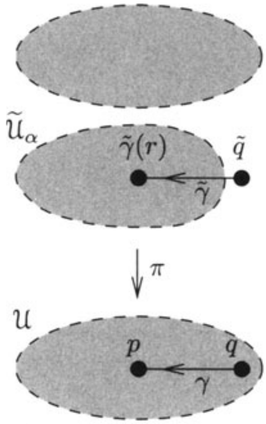


FIGURE 11.3. Proof that $\pi^{-1}(\mathcal{U}) \subset \bigcup_\alpha \widetilde{\mathcal{U}}_\alpha$.

It remains only to show that $\pi\colon \widetilde{\mathcal{U}}_\alpha \to \mathcal{U}$ is a diffeomorphism for each α. It is certainly a local diffeomorphism (because π is). It is bijective because its inverse can be constructed explicitly: it is the map sending each radial geodesic starting at p to its lift starting at $\tilde{p}_\alpha$. This completes the proof. □

Because of this theorem, a complete, simply-connected Riemannian manifold with nonpositive sectional curvature is called a *Cartan–Hadamard manifold.* An immediate consequence of the Cartan–Hadamard theorem is that there are stringent topological restrictions on which manifolds can carry metrics of nonpositive sectional curvature. For example, if M is a product of compact manifolds $M_1 \times M_2$ where either M_1 or M_2 is simply connected (such as, for example, $S^1 \times S^2$), then any metric on M must have positive sectional curvature somewhere. With a little algebraic topology, one can obtain more information: for example, any manifold whose universal cover is contractible is *aspherical,* which means that the higher homotopy groups $\pi_k(M)$ vanish for $k > 1$ (see [Whi78]), so many manifolds cannot admit metrics of nonpositive curvature.

Manifolds of Positive Curvature

Next we consider manifolds with positive sectional curvature. Our comparison theorems do not tell us anything about manifolds whose curvature is bounded *below* instead of above. Nevertheless, clever analysis of the index form can still lead to significant conclusions, as the proof of the following theorem shows. We need one definition: the *diameter* of a Riemannian

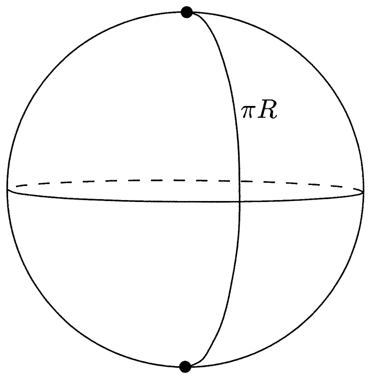


FIGURE 11.4. The diameter of S^n_R is πR.

manifold is

$$\operatorname{diam}(M) := \sup\{d(p,q) : p,q \in M\}.$$

Note that the diameter of the round sphere of radius R is πR (not $2R$), since the Riemannian distance between antipodal points is πR (Figure 11.4).

Theorem 11.7. (Bonnet's Theorem) *Let M be a complete, connected Riemannian manifold all of whose sectional curvatures are bounded below by a positive constant $1/R^2$. Then M is compact, with a finite fundamental group, and with diameter less than or equal to πR.*

Proof. The first step is to show that the diameter of M is no greater than πR. Suppose the contrary: then there are points $p, q \in M$, and (by the Hopf–Rinow theorem) a minimizing unit speed geodesic segment γ from p to q of length $L > \pi R$. Since γ is minimizing, its index form is nonnegative. We will derive a contradiction by constructing a proper normal vector field V along γ such that $I(V,V) < 0$.

Let E be any parallel normal unit vector field along γ, and let

$$V(t) = \left(\sin \frac{\pi t}{L}\right) E(t).$$

Observe that V vanishes at $t = 0$ and $t = L$, so V is a proper normal vector field along γ. (Note the similarity between V and the formulas (10.5), (10.6) for Jacobi fields on the sphere of radius L/π.) By direct computation,

$$D_t V(t) = \frac{\pi}{L}\left(\cos \frac{\pi t}{L}\right) E(t),$$

$$D_t^2 V(t) = -\frac{\pi^2}{L^2}\left(\sin \frac{\pi t}{L}\right) E(t),$$

and so

$$\begin{aligned} I(V,V) &= -\int_0^L \langle D_t^2 V + R(V,\dot\gamma)\dot\gamma, V\rangle\, dt \\ &= \int_0^L \left\langle \frac{\pi^2}{L^2}\Big(\sin\frac{\pi t}{L}\Big)E - \Big(\sin\frac{\pi t}{L}\Big)R(E,\dot\gamma)\dot\gamma, \Big(\sin\frac{\pi t}{L}\Big)E\right\rangle dt \\ &= \int_0^L \Big(\sin^2\frac{\pi t}{L}\Big)\Big(\frac{\pi^2}{L^2} - Rm(E,\dot\gamma,\dot\gamma,E)\Big)\, dt. \end{aligned}$$

Since E and $\dot\gamma$ are orthonormal, $Rm(E,\dot\gamma,\dot\gamma,E)$ is equal to the sectional curvature of the plane they span, and so our estimate on sectional curvature gives

$$I(V,V) \le \int_0^L \Big(\sin^2\frac{\pi t}{L}\Big)\Big(\frac{\pi^2}{L^2} - \frac{1}{R^2}\Big)\, dt < 0.$$

Therefore our geodesic of length $L > \pi R$ cannot be minimizing, so the diameter of M is at most πR.

To show that M is compact, we just choose a basepoint p and note that every point in M can be connected to p by a geodesic segment of length at most πR. Therefore, $\exp_p\colon \overline{B}_{\pi R}(0) \to M$ is surjective, so M is the continuous image of a compact set.

Finally, let $\pi\colon \widetilde M \to M$ denote the universal covering space of M, with the metric $\tilde g := \pi^* g$. $\widetilde M$ is complete by the result of Problem 6-11, and $\tilde g$ also has sectional curvatures bounded below by $1/R^2$, so $\widetilde M$ is compact by the argument above. By the theory of covering spaces (see [Mas67, Corollary V.7.5]), there is a one-to-one correspondence between $\pi_1(M)$ and the inverse image $\pi^{-1}(p)$ of any point $p \in M$. If $\pi_1(M)$ were infinite, therefore, $\pi^{-1}(p)$ would be an infinite discrete set in $\widetilde M$, contradicting the compactness of $\widetilde M$. Thus $\pi_1(M)$ is finite. □

It is rather surprising that the conclusions of Bonnet's theorem hold with the much weaker assumption of strictly positive Ricci tensor, as the following theorem shows.

Theorem 11.8. (Myers's Theorem) *Suppose M is a complete, connected Riemannian n-manifold whose Ricci tensor satisfies the following inequality for all $V \in TM$:*

$$Rc(V,V) \ge \frac{n-1}{R^2}|V|^2.$$

Then M is compact, with a finite fundamental group, and diameter at most πR.

Proof. As in the proof of Bonnet's theorem, it suffices to prove the diameter estimate. As before, let γ be a minimizing unit speed geodesic segment of

length $L > \pi R$. Let $(E_1, \dots, E_n)$ be a parallel orthonormal frame along γ such that $E_n = \dot\gamma$, and for each $i = 1, \dots, n-1$ let V_i be the proper normal vector field

$$V_i(t) = \left(\sin\frac{\pi t}{L}\right)E_i(t).$$

By the same computation as before,

$$I(V_i, V_i) = \int_0^L \left(\sin^2\frac{\pi t}{L}\right)\left(\frac{\pi^2}{L^2} - Rm(E_i, \dot\gamma, \dot\gamma, E_i)\right) dt. \tag{11.1}$$

In this case, we cannot conclude that each of these terms is negative. However, because $\{E_i\}$ is an orthonormal frame, the Ricci tensor at points along γ is given by

$$Rc(\dot\gamma, \dot\gamma) = \sum_{i=1}^{n} Rm(E_i, \dot\gamma, \dot\gamma, E_i) = \sum_{i=1}^{n-1} Rm(E_i, \dot\gamma, \dot\gamma, E_i)$$

(because $Rm(E_n, \dot\gamma, \dot\gamma, E_n) = Rm(\dot\gamma, \dot\gamma, \dot\gamma, \dot\gamma) = 0$). Therefore, summing (11.1) over i gives

$$\begin{aligned}\sum_{i=1}^{n-1} I(V_i, V_i) &= \int_0^L \left(\sin^2\frac{\pi t}{L}\right)\left((n-1)\frac{\pi^2}{L^2} - Rc(\dot\gamma, \dot\gamma)\right) dt \\ &\leq \int_0^L \left(\sin^2\frac{\pi t}{L}\right)\left(\frac{(n-1)\pi^2}{L^2} - \frac{n-1}{R^2}\right) dt < 0.\end{aligned}$$

This means at least one of the terms $I(V_i, V_i)$ must be negative, and again we have a contradiction to γ being minimizing. □

One of the most useful applications of Myers's theorem is to Einstein metrics. If g is a complete Einstein metric with positive scalar curvature, then $Rc = \frac{1}{n}Sg$ satisfies the hypotheses of the theorem; it follows that complete, noncompact Einstein manifolds must have nonpositive scalar curvature. On the other hand, it is possible for complete, noncompact manifolds to have strictly positive Ricci or even sectional curvature, as long as it gets arbitrarily close to zero, as the following example shows.

Exercise 11.1. Let $M \subset \mathbf{R}^{n+1}$ be the paraboloid $\{(x^1, \dots, x^n, y) : y = |x|^2\}$ with the induced metric (see Problem 8-2). Show that M has strictly positive sectional curvature everywhere.

There is much more that can be said in the case of positive sectional curvature, using more elaborate versions of the methods we have developed here. One question you might already have asked yourself is whether there are analogues of our comparison theorems when the curvature is bounded

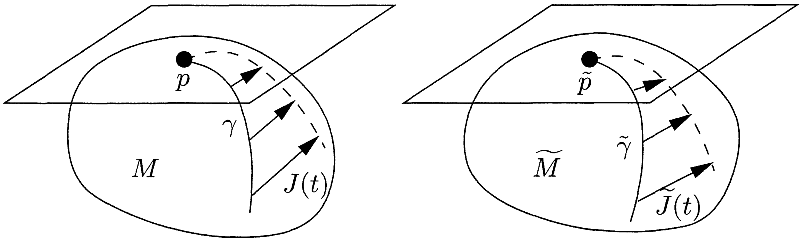


FIGURE 11.5. Setup for the Rauch comparison theorem.

below instead of above. Our proof of the Jacobi field comparison theorem definitely does not work if a lower bound on curvature is substituted for the upper bound, because the step involving the Schwartz inequality is not reversible (except in dimension 2—see Problem 11-1).

Nonetheless, the analogues of all three of these results are true with curvature bounded above, but the proofs are considerably more involved. The key fact is the following very general comparison theorem. The proof would take us too far afield, so we state it without proof.

Theorem 11.9. (Rauch Comparison Theorem) *Let M and $\widetilde{M}$ be Riemannian manifolds, let $\gamma\colon [0,T] \to M$ and $\tilde\gamma\colon [0,T] \to \widetilde{M}$ be unit speed geodesic segments such that $\tilde\gamma(0)$ has no conjugate points along $\tilde\gamma$, and let J, $\widetilde{J}$ be normal Jacobi fields along γ and $\tilde\gamma$ such that $J(0) = \widetilde{J}(0) = 0$ and $|D_t J(0)| = |\widetilde{D}_t \widetilde{J}(0)|$ (Figure 11.5). Suppose that the sectional curvatures of M and $\widetilde{M}$ satisfy $K(\Pi) \le \widetilde{K}(\widetilde{\Pi})$ whenever $\Pi \subset T_{\gamma(t)}M$ is a 2-plane containing $\dot\gamma(t)$ and $\widetilde{\Pi} \subset T_{\tilde\gamma(t)}\widetilde{M}$ is a 2-plane containing $\dot{\tilde\gamma}(t)$. Then $|J(t)| \ge |\widetilde{J}(t)|$ for all $t \in [0,T]$.*

You can find proofs in [dC92], [CE75], and [Spi79, volume 4]. Letting $\widetilde{M}$ be one of our constant curvature model spaces, we recover the Jacobi field comparison theorem above. On the other hand, if instead we take M to have constant curvature, we get the same result with the inequalities reversed.

The most successful applications of the Rauch comparison theorem have been to prove "pinching theorems." A manifold is said to be *δ-pinched* if all sectional curvatures satisfy

$$\delta \frac{1}{R^2} \le K(\Pi) \le \frac{1}{R^2},$$

for some $\delta, R > 0$, and *strictly δ-pinched* if the first inequality is strict. The following celebrated theorem was originally proved by Marcel Berger and Walter Klingenberg in the early 1960s.

Theorem 11.10. (The Sphere Theorem) *Suppose M is a complete, simply-connected, Riemannian n-manifold that is strictly $\frac{1}{4}$-pinched. Then M is homeomorphic to $\mathbf{S}^n$.*

The proof, which can be found in [CE75] or [dC92], is an elaborate application of the Rauch comparison theorem together with the Morse index theorem mentioned in Chapter 10. This result is sharp, at least in even dimensions, because the Fubini–Study metrics on complex projective spaces are $\frac{1}{4}$-pinched (Problem 8-12).

Using techniques of partial differential equations can lead to even stronger conclusions in some cases. For instance, in 1982, Richard Hamilton [Ham82] proved the following very striking result on 3-manifolds.

Theorem 11.11. (Hamilton) *Suppose M is a simply-connected compact Riemannian 3-manifold with strictly positive Ricci curvature. Then M is diffeomorphic to $\mathbf{S}^3$.*

Manifolds of Constant Curvature

Our last application of Jacobi field techniques is to give a global characterization of complete manifolds of constant sectional curvature.

Theorem 11.12. (Uniqueness of Constant Curvature Metrics) *Let M be a complete, simply-connected Riemannian n-manifold with constant sectional curvature C. Then M is isometric to one of the model spaces $\mathbf{R}^n$, $\mathbf{S}^n_R$, or $\mathbf{H}^n_R$.*

Proof. It is easiest to handle the cases of positive and nonpositive sectional curvature separately. First suppose $C \leq 0$. Then the Cartan–Hadamard theorem says that for any $p \in M$, $\exp_p : T_pM \to M$ is a covering map. Since M is simply connected, it is a diffeomorphism. The pulled-back metric $\tilde{g} := \exp_p^* g$, therefore, is a globally defined metric on T_pM with constant sectional curvature C, and $\exp_p : (T_pM, \tilde{g}) \to (M, g)$ is a global isometry. Moreover, since Euclidean coordinates for T_pM are normal coordinates for $\tilde{g}$, it must be given by one of the cases of formula (10.8); these in turn are globally isometric to $\mathbf{R}^n$ if $C = 0$ and $\mathbf{H}^n_R$ if $C = -1/R^2$.

In case $C = 1/R^2 > 0$, we have to argue a little differently. Let $\{N, -N\}$ be the north and south poles in $\mathbf{S}^n_R$, and observe that $\exp_N$ is a diffeomorphism from $B_{\pi R}(0) \subset T_N\mathbf{S}^n_R$ to $\mathbf{S}^n_R - \{-N\}$. On the other hand, choosing any point $p \in M$, the conjugate point comparison theorem shows that p has no conjugate points closer than πR, so $\exp_p$ is at least a local diffeomorphism on $B_{\pi R}(0) \subset T_pM$. If we choose any linear isometry $\varphi : T_N\mathbf{S}^n_R \to T_pM$ (Figure 11.6), then $(\exp_p \circ \varphi)^* g$ and $\exp_N^* \mathring{g}_R$ are both metrics of constant curvature $1/R^2$ on $B_R(0) \subset T_N\mathbf{S}^n_R$, and Euclidean coordinates on $T_N S^n_R$ are normal coordinates for both (since the radial line

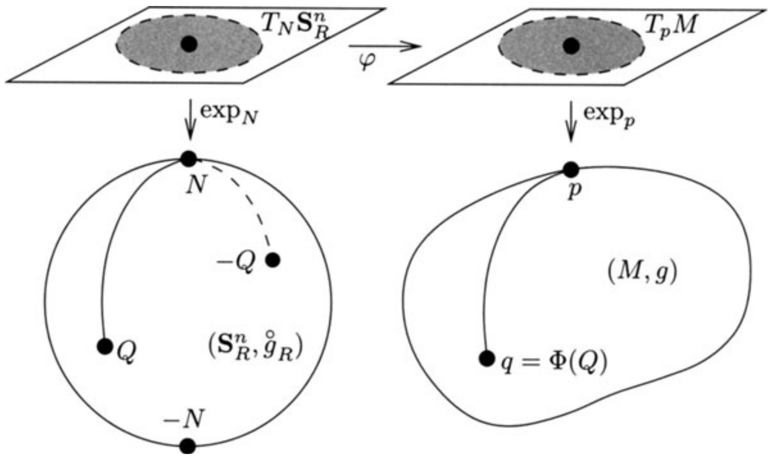


FIGURE 11.6. Constructing an isometry when $C > 0$.

segments are geodesics). Therefore, Proposition 10.9 shows that they are equal, so the map $\Phi\colon \mathbf{S}^n_R - \{-N\} \to M$ given by $\Phi = \exp_p \circ \varphi \circ \exp_N^{-1}$ is a local isometry.

Now choose any point $Q \in \mathbf{S}^n_R$ other than N or $-N$, and let $q = \Phi(Q) \in M$. Using the isometry $\widetilde{\varphi} = \Phi_*\colon T_Q\mathbf{S}^n_R \to T_qM$, we can construct a similar map $\widetilde{\Phi} = \exp_q \circ \widetilde{\varphi} \circ \exp_Q^{-1}\colon \mathbf{S}^n_R - \{-Q\} \to M$, and the same argument shows that $\widetilde{\Phi}$ is a local isometry. Because $\Phi(Q) = \widetilde{\Phi}(Q)$ and $\Phi_* = \widetilde{\Phi}_*$ at Q by construction, Φ and $\widetilde{\Phi}$ must agree where they overlap by Problem 5-7. Putting them together, therefore, we get a globally defined local isometry $F\colon \mathbf{S}^n_R \to M$. After noting that M is compact by Bonnet's theorem, we complete the proof by appealing to Exercise 11.2 below. □

Exercise 11.2. Show that any local diffeomorphism between compact, connected manifolds is a covering map.

Theorem 11.12 is a special case of a rather more complicated result, the Cartan–Ambrose–Hicks theorem, which says roughly that two simply-connected manifolds, all of whose sectional curvatures at corresponding points are equal to each other, must be isometric. The main idea of the proof is very similar to what we have done here; the trick is in making precise sense of the notion of "corresponding points," and of what it means for nonconstant sectional curvatures to be equal at different points of different manifolds. See [Wol84] or [CE75] for the complete statement and proof.

Combining our classification of simply-connected manifolds of constant curvature with the characterization of their isometry groups given in Prob-

lem 5-8, we obtain finally the following description of all complete manifolds of constant curvature.

Corollary 11.13. (Classification of Constant Curvature Metrics) *Suppose M is a complete, connected Riemannian manifold with constant sectional curvature. Then M is isometric to $\widetilde{M}/\Gamma$, where $\widetilde{M}$ is one of the constant curvature model spaces $\mathbf{R}^n$, $\mathbf{S}^n_R$, or $\mathbf{H}^n_R$, and Γ is a discrete subgroup of $\mathfrak{I}(\widetilde{M})$, isomorphic to $\pi_1(M)$, and acting freely and properly discontinuously on $\widetilde{M}$.*

Proof. If $\pi\colon \widetilde{M} \to M$ is the universal covering space of M with the lifted metric $\tilde{g} = \pi^* g$, the preceding theorem shows that $(\widetilde{M}, \tilde{g})$ is isometric to one of the model spaces. From covering space theory [Sie92, Mas67] it follows that the group Γ of covering transformations is isomorphic to $\pi_1(M)$ and acts freely and properly discontinuously on $\widetilde{M}$, and M is diffeomorphic to the quotient $\widetilde{M}/\Gamma$. Moreover, if φ is any covering transformation, $\pi \circ \varphi = \pi$, and so $\varphi^* \tilde{g} = \varphi^* \pi^* g = \pi^* g = \tilde{g}$, so Γ acts by isometries. Finally, suppose $\{\varphi_i\} \subset \Gamma$ is an infinite set with an accumulation point in $\mathfrak{I}(\widetilde{M})$. Since the action of Γ is fixed-point free, for any point $\tilde{p} \in \widetilde{M}$ the set $\{\varphi_i(\tilde{p})\}$ is infinite, and by continuity of the action it has an accumulation point in $\widetilde{M}$. But this is impossible, since the points $\{\varphi_i(\tilde{p})\}$ all project to the same point in M, and so form a discrete set. Thus Γ is discrete in $\mathfrak{I}(\widetilde{M})$. □

A complete, connected Riemannian manifold with constant sectional curvature is called a *space form.* This result essentially reduces the classification of space forms to group theory. Nevertheless, the group-theoretic problem is still far from easy.

The spherical space forms were classified in 1972 by Joseph Wolf [Wol84]; the proof is intimately connected with the representation theory of finite groups. Although the only 2-dimensional ones are the sphere and the projective plane, already in dimension 3 there are many interesting examples. Some notable ones are the *lens spaces* obtained as quotients of $\mathbf{S}^3 \subset \mathbf{C}^2$ by cyclic groups rotating the two complex coordinates through different angles; and the quotients of $SO(3)$ (which is diffeomorphic to $\mathbf{RP}^3$ and is therefore already a quotient of $\mathbf{S}^3$) by the *dihedral groups,* the symmetry groups of regular 3-dimensional polyhedra.

The complete classification of Euclidean space forms is known only in low dimensions. For example, there are 10 classes of nondiffeomorphic compact Euclidean space forms of dimension 3, and 75 classes in dimension 4. The fundamental groups of compact Euclidean space forms are examples of *crystallographic groups,* which are discrete groups of Euclidean isometries with compact quotients, and which have been studied extensively by physicists as well as geometers. (A quotient of $\mathbf{R}^n$ by a crystallographic group is a space form provided it is a manifold, which is true whenever the crystallographic group has no elements of finite order.) It is known in general

that Euclidean space forms are quotients of flat tori, but the classification in higher dimensions is still elusive. See [Wol84] for a complete survey of the state of the art as of 1972.

Finally, the study of hyperbolic space forms is a vast and rich subject, the surface of which has barely been scratched.

Problems

11-1. Show that when the dimension of M is 2, the argument of Theorem 11.2 can be adapted to give an explicit upper bound for $|J(t)|$ when the Gaussian curvature is bounded below.

11-2. Adapt the argument of Theorem 11.1 to prove the following generalizations of the Sturm comparison theorem, also due to Sturm.

(a) Suppose a, b are continuous functions on an open interval I with $a \geq b$, and u, v are nontrivial solutions to

$$\ddot{u}(t) + a(t)u(t) = 0$$
$$\ddot{v}(t) + b(t)v(t) = 0$$

on I. Then between any two zeros of v there must be at least one zero of u, unless $a \equiv b$ and u and v are constant multiples of each other.

(b) **(Sturm Separation Theorem)** Suppose a is continuous on an interval I, and u_1, u_2 are two linearly independent solutions on I to

$$\ddot{u}(t) + a(t)u(t) = 0.$$

Show that the zeros of u_1 and u_2 are strictly alternating.

11-3. Suppose M is a Cartan–Hadamard manifold whose sectional curvature is bounded above by a negative constant C. Show that the volume of any geodesic ball in M at least as large as that of the ball with the same radius in hyperbolic space of curvature C.

References

[AP94] Marco Abate and Giorgio Patrizio. *Finsler Metrics, a Global Approach: with Applications to Geometric Function Theory.* Springer-Verlag, New York, 1994.

[Bes87] Arthur L. Besse. *Einstein Manifolds.* Springer-Verlag, Berlin, 1987.

[Boo86] William M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press, Orlando, second edition, 1986.

[BR78] Garrett Birkhoff and Gian-Carlo Rota. *Ordinary Differential Equations.* Wiley, New York, third edition, 1978.

[CE75] Jeff Cheeger and David G. Ebin. *Comparison Theorems in Riemannian Geometry.* North-Holland, Amsterdam, 1975.

[Cha93] Isaac Chavel. *Riemannian Geometry: A Modern Introduction.* Cambridge University Press, Cambridge, 1993.

[Che96] Shiing-Shen Chern. Finsler geometry is just Riemannian geometry without the quadratic restriction. *Notices Amer. Math. Soc.*, 43:959–963, 1996.

[dC92] Manfredo Perdigão do Carmo. *Riemannian Geometry.* Birkhäuser, Boston, 1992.

[Gau65] Carl Friedrich Gauss. *General Investigations of Curved Surfaces.* Raven Press, New York, 1965.

[GHL87] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian Geometry.* Springer-Verlag, Berlin, 1987.

[Ham82] Richard Hamilton. Three-manifolds with positive Ricci curvature. *J. Differential Geom.*, 17:255–306, 1982.

[HE73] S. W. Hawking and G. F. R. Ellis. *The Large-Scale Structure of Space-Time.* Cambridge University Press, Cambridge, 1973.

[Hop35] Heinz Hopf. Über die Drehung der Tangenten und Sehnen ebener Kurven. *Compositio Math.*, 2:50–62, 1935.

[Hop83] Heinz Hopf. *Differential Geometry in the Large.* Number 1000 in Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1983.

[JP93] Marek Jarnicki and Peter Pflug. *Invariant Distances and Metrics in Complex Analysis.* Walter de Gruyter and Co., Berlin, 1993.

[Kaz85] Jerry Kazdan. *Prescribing the Curvature of a Riemannian Manifold.* Number 57 in CBMS Regional Conf. Ser. in Math. Amer. Math. Soc., Providence, 1985.

[KN63] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry*, volume I–II. Interscience, New York, 1963.

[Kob72] Shoshichi Kobayashi. *Transformation Groups in Differential Geometry.* Springer-Verlag, Berlin, 1972.

[Law80] H. Blaine Lawson, Jr. *Lectures on Minimal Submanifolds.* Publish or Perish, Berkeley, 1980.

[LP87] John M. Lee and Thomas H. Parker. The Yamabe problem. *Bull. Amer. Math. Soc. (N.S.)*, 17:37–91, 1987.

[Mas67] William S. Massey. *Algebraic Topology: An Introduction.* Springer-Verlag, Berlin, 1967.

[Mor93] Frank Morgan. *Riemannian Geometry: A Beginner's Guide.* Jones and Bartlett, Boston, 1993.

[Nas56] John Nash. The imbedding problem for Riemannian manifolds. *Ann. Math.*, 63:20–63, 1956.

[O'N83] Barrett O'Neill. *Semi-Riemannian Geometry with Applications to General Relativity.* Academic Press, New York, 1983.

[Rad25] Tibor Radó. Über den Begriff der Riemannschen Fläche. *Acta Litt. Sci. Szeged.*, 2:101–121, 1925.

[Sie92] Allan J. Sieradski. *An Introduction to Topology and Homotopy.* PWS-Kent, Boston, 1992.

[Spi79] Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume I–V. Publish or Perish, Berkeley, 1979.

[Str86] Robert S. Strichartz. Sub-Riemannian geometry. *J. Differential Geom.*, 24:221–263, 1986.

[Whi78] George W. Whitehead. *Elements of Homotopy Theory.* Springer-Verlag, Berlin, 1978.

[Wol84] Joseph A. Wolf. *Spaces of Constant Curvature.* Publish or Perish, Berkeley, fifth edition, 1984.

Index

Graduate Texts in Mathematics

continued from page ii

119 ROTMAN. An Introduction to Algebraic Topology.
120 ZIEMER. Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.
121 LANG. Cyclotomic Fields I and II. Combined 2nd ed.
122 REMMERT. Theory of Complex Functions. *Readings in Mathematics*
123 EBBINGHAUS/HERMES et al. Numbers. *Readings in Mathematics*
124 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part III.
125 BERENSTEIN/GAY. Complex Variables: An Introduction.
126 BOREL. Linear Algebraic Groups.
127 MASSEY. A Basic Course in Algebraic Topology.
128 RAUCH. Partial Differential Equations.
129 FULTON/HARRIS. Representation Theory: A First Course. *Readings in Mathematics*
130 DODSON/POSTON. Tensor Geometry.
131 LAM. A First Course in Noncommutative Rings.
132 BEARDON. Iteration of Rational Functions.
133 HARRIS. Algebraic Geometry: A First Course.
134 ROMAN. Coding and Information Theory.
135 ROMAN. Advanced Linear Algebra.
136 ADKINS/WEINTRAUB. Algebra: An Approach via Module Theory.
137 AXLER/BOURDON/RAMEY. Harmonic Function Theory.
138 COHEN. A Course in Computational Algebraic Number Theory.
139 BREDON. Topology and Geometry.
140 AUBIN. Optima and Equilibria. An Introduction to Nonlinear Analysis.
141 BECKER/WEISPFENNING/KREDEL. Gröbner Bases. A Computational Approach to Commutative Algebra.
142 LANG. Real and Functional Analysis. 3rd ed.
143 DOOB. Measure Theory.
144 DENNIS/FARB. Noncommutative Algebra.
145 VICK. Homology Theory. An Introduction to Algebraic Topology. 2nd ed.
146 BRIDGES. Computability: A Mathematical Sketchbook.
147 ROSENBERG. Algebraic *K*-Theory and Its Applications.
148 ROTMAN. An Introduction to the Theory of Groups. 4th ed.
149 RATCLIFFE. Foundations of Hyperbolic Manifolds.
150 EISENBUD. Commutative Algebra with a View Toward Algebraic Geometry.
151 SILVERMAN. Advanced Topics in the Arithmetic of Elliptic Curves.
152 ZIEGLER. Lectures on Polytopes.
153 FULTON. Algebraic Topology: A First Course.
154 BROWN/PEARCY. An Introduction to Analysis.
155 KASSEL. Quantum Groups.
156 KECHRIS. Classical Descriptive Set Theory.
157 MALLIAVIN. Integration and Probability.
158 ROMAN. Field Theory.
159 CONWAY. Functions of One Complex Variable II.
160 LANG. Differential and Riemannian Manifolds.
161 BORWEIN/ERDÉLYI. Polynomials and Polynomial Inequalities.
162 ALPERIN/BELL. Groups and Representations.
163 DIXON/MORTIMER. Permutation Groups.
164 NATHANSON. Additive Number Theory: The Classical Bases.
165 NATHANSON. Additive Number Theory: Inverse Problems and the Geometry of Sumsets.
166 SHARPE. Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.
167 MORANDI. Field and Galois Theory.
168 EWALD. Combinatorial Convexity and Algebraic Geometry.
169 BHATIA. Matrix Analysis.
170 BREDON. Sheaf Theory. 2nd ed.
171 PETERSEN. Riemannian Geometry.
172 REMMERT. Classical Topics in Complex Function Theory.
173 DIESTEL. Graph Theory.
174 BRIDGES. Foundations of Real and Abstract Analysis.
175 LICKORISH. An Introduction to Knot Theory.
176 LEE. Riemannian Manifolds.